Evaluation of deep pose detectors for automatic analysis of film style

Hui-Yin Wu, Luan Nguyen, Yoldoz Tabei, Lucile Sassatelli

Eurographics Workshop on Intelligent Cinematography and Editing (2022)
R. Ronfard and H.-Y. Wu (Editors)

Evaluation of deep pose detectors for automatic analysis of film style

Hui-Yin Wu[1], Luan Nguyen[2], Yoldoz Tabei[2] and Lucile Sassatelli[2,3]

[1]Université Côte d'Azur, Inria, France
[2]Université Côte d'Azur, CNRS, I3S, France
[3]Institut Universitaire de France

Abstract

Identifying human characters and how they are portrayed on-screen is inherently linked to how we perceive and interpret the story and artistic value of visual media. Building computational models sensible towards story will thus require a formal representation of the character. Yet this kind of data is complex and tedious to annotate on a large scale. Human pose estimation (HPE) can facilitate this task, to identify features such as position, size, and movement that can be transformed into input to machine learning models, and enable higher artistic and storytelling interpretation. However, current HPE methods operate mainly on non-professional image content, with no comprehensive evaluation of their performance on artistic film.
Our goal in this paper is thus to evaluate the performance of HPE methods on artistic film content. We first propose a formal representation of the character based on cinematography theory, then sample and annotate 2700 images from three datasets with this representation, one of which we introduce to the community. An in-depth analysis is then conducted to measure the general performance of two recent HPE methods on metrics of precision and recall for character detection , and to examine the impact of cinematographic style. From these findings, we highlight the advantages of HPE for automated film analysis, and propose future directions to improve their performance on artistic film content.

CCS Concepts
• ***Computing methodologies*** → ***Computer vision; Neural networks;*** • ***Applied computing*** → ***Media arts;***

1. Introduction

Our imaginations of gender, society, and identity are strongly influenced by the film content that we consume every day – in movie theaters, on TV, and online platforms. As the technical bar to creating visual content lowers, the challenge to analyze and understand the constructs that are communicated in our media grows, calling for automated approaches that can address this quantitative data challenge, and simultaneously enrich the qualitative analysis.

Amongst various visual features of film, character representation is central to understanding story and plot development. Cinematography has long established practices to portray characters through framing properties such as size and angle. But how can we conduct a quantitative analysis on character?

Deep human pose detectors have an unprecedented advantage to the gross extraction of character from image data. However, the most well-known and high-performance methods have been trained on mostly non-artistic and non-professional content. Their performance on film data with complex visual arrangements has not yet been evaluated. An initial exploration by running human pose estimation (HPE) on a few film clips will quickly reveal a large number of detection errors and limitations, but the extent of these limitations has not yet been fully examined.

In this paper we address this question with three contributions:

- We propose a formal representation of on-screen character based on film practice that can potentially be automated by Human Pose Estimation (HPE) methods, to progress towards automatic analysis of film character representation, presented in Section 3.
- We develop a framework to evaluate HPE methods on a variety of datasets. This involves the review and selection of relevant film and non-film datasets, sampling of datasets to ensure a wide variety of images, and the development of annotation tools to label cinematographic features of character representation and pose estimation quality. We consider a non-film and 2 film datasets, one of which we introduce to the community. We hence obtain over 2700 images from three datasets, annotated with cinematographic features and human pose ground truth. This is presented in detail in Section 4.3.
- We assess the performance in character detection of HPE (trained on non-film data) on the datasets. The analysis of results disaggregated over groups of frames with common cinematographic features allows us to quantitatively pinpoint the character representations where current HPE methods fail. We thereby identify new types of data to be annotated and inserted in training sets to improve HPE on film data. This is presented in Section 5.4.

2. Related work

Here we first review the domain of HPE with categories of representative deep learning approaches and remaining challenges. Second, we discuss character representation with formal constructs from film theory. Third, we present existing datasets to progress towards automating the analysis of film character representation.

2.1. Human Pose Estimation

We consider the problem of Human Pose Estimation (HPE) in 2D, which consists in estimating the 2D positions of human body parts from an input image. It involves localizing keypoints of the body (joints such as shoulders, wrists, hips, ankle, etc. and possible face or hand landmarks), and connecting those to obtain a skeleton for each person in the image. Existing methods, now based on Deep Learning for best performance, can be categorized into top-down and bottom-up approaches. Top-down methods consist in first detecting all people in the image, then predicting keypoint locations for each person. In contrast, bottom-up methods first localize keypoints, then group keypoints to obtain the individual. Numerous methods in each category exist. A recent and high-performing representative of top-down approaches is High-Resolution Net (HRNet) presented by Sun et al. [SXLW19], producing high-resolution heatmap of keypoints to improve localization. A major representative of bottom-up approaches is OpenPose, introduced by Cao et al. in [CSWS17]. OpenPose first predicts keypoints coordinates as well as Part Affinity Fields encoding the orientation of the limbs, which are then used to solve a relaxed version of the matching keypoints problem. Recently, Geng et al. introduced Disentangled Keypoint Regression (DEKR) in [GSX*21]. DEKR is a bottom-up approach which processes the features learned by an external backbone, taken as HRNet, to regress offset maps centered on each pixel for every keypoint, and has been shown to outperform HRNet. Bottom-up approaches are comparatively less computationally intensive, with OpenPose and DEKR reaching real-time performance, while top-down approaches tend to produce more accurate results. A more comprehensive survey can be found in [ZWY*20]. In videos, one can leverage temporal information to improve pose estimation [ZWCS20, BLC*21]. Pose estimation through time and pose tracking is a relatively new topic. In this article, we only consider image-based methods.

HPE from images still face significant challenges, including reliable detection under body occlusion, and limited data for rare poses and angles. It is therefore important to develop techniques to automatically augment existing annotated datasets, but also to select complementary data to be manually annotated for keypoints localization for domain-specific tasks such as film style analysis. That is why is this article, we analyze specifically the challenges that cinematic content poses to existing HPE methods.

HPE evaluation for cinematography : There are several ways to evaluate HPE performance in the context of cinematography. For example, one could attempt to predict the position, angle or size of each actor from the HPE results, and compare them with ground truth. Also, two HPE methods could be compared on the same non-annotated dataset by identifying the number of manual corrections to make after each method has estimated poses. In this article, we present a first principled evaluation of HPE methods on character detection depending on the cinematographic features of the image.

2.2. Film character analysis

The computational analysis of character in films has long drawn interest. Devoted vocabulary such as the Prose Storyboard language [RVB13] and Film Editing Patterns [WPRC18] have been developed from film textbooks and practice [Mas65, Zet16]. Descriptions with these languages are centered on visual features surrounding the character, such as position, angle, and size. In particular, our analysis in this paper is based on vocabulary from Film Editing Patterns [WPRC18], for which the constraints on these features can be expressed and solved for sequences of one or more shots. However, there exists a wide gap between the amount of cinematographic knowledge these vocabulary can express, and the features that can be extracted from film clips. At the moment, one must rely heavily on human-annotated datasets to have sufficient features matching the stylistic patterns we wish to analyze, calling out the need for more automated approaches to have larger datasets with rich representations [WGLC17].

Recent encouraging work by Somandepalli et al. [SGM*21] has demonstrated the power of multi-modal media analysis tools involving face detection, audio, and script to automate analysis of character representation. They identify character interactions (i.e., when characters appear together in a film) and quantify gender disparities by measuring female screen time and speaking time over 600 films. Courant et al. [CLCK21] have also integrated pose detection as one of the high-level features for character understanding, in addition to other camera and framing characteristics. However, to date, no works have used full human poses as an integral component of a film analysis framework to extract cinematographic features of a clip such as shot size, angle, body part visibility, or artistic framings (e.g., over the shoulder, cowboy shots).

2.3. Datasets

The availablity and quality of human pose datasets is equally crucial to this work. The Microsoft Common Objects in Context (COCO) [LMB*14] and Max Planck Institute for Informatics (MPII) Human Pose [APGS14] are the reference datasets for 2D HPE. MPII contains 25K images extracted from YouTube videos of everyday human activities, annotated with 16 joints. COCO is a large-scale dataset that has played an instrumental role in training models for tasks such as object detection and classification, semantic and instance segmentation, and keypoint detection. COCO2017 contains 106K images obtained from the Web and with (most) main individuals annotated with 17 joints.

The only dataset of film data annotated with keypoints is the Frames Labeled In Cinema (FLIC) dataset [ST13], containing 20K images obtained from 30 Hollywood movies. The FLIC-plus version contains a subset of 17K images without scene overlap between train and test images. In the FLIC datasets, only 10 upper-body joints have been annotated, which is limiting if one wants to investigate how characters' bodies are represented (and possibly objectified) on screen.

Other movie datasets exist, but do not contain body keypoint annotations. For example, the MovieNet dataset has been introduced by Huang et al. in [HXR*20]. MovieNet is made of multimodal data obtained from 1100 movies. It contains annotations for 1.1M characters' bounding boxes and identities, tags for places, actions, cinematic styles, and scripts.

3. Overview: character representation and pose

We propose an analysis of human pose estimators based on performance metrics associated with specific frame criteria: (1) precision, recall, and pose keypoint accuracy measures that have been used for pre-existing benchmarks, and (2) a set of cinematographic labels extending Film Editing Patterns [WPRC18] focused on character representation, and spanning six large categories: character size, character angle (both pitch and yaw), on-screen position, number of characters, body part visibility, and artistic shots. Here is a brief description of each category:

- **Character size**: the size of the character on screen based on the relative size to the screen,
- **Character angle**: the angle of the camera, both in vertical pitch and horizontal yaw of the character,
- **On-screen position**: the horizontal and vertical position of the point at the center of the character's two eyes on a 3x3 grid. Empty when the head is not visible,
- **Body part visibility**: shots that contain only part of the body, or with certain body parts hidden,
- **Artistic Style**: numerous typical framings used in films shown in Figure 1, non-standing actions (lying, sitting, acrobatic), and humanoid-like objects (dummy).

When multiple characters are present in the scene, we evaluate the characteristics for the three largest characters on screen. The full list of labels for each category is shown in Table 1.

Table 1: *Labels for six shot categories evaluated. Shot sizes use abbreviations (V:Very, L:Long, S:Shot, M:Medium, CU:Closeup, X:Extreme). Artistic framings are illustrated in Figure 1*

Category	Labels
char. size	VLS, LS, MLS, MS, MCU, CU, XCU
character angle	*pitch*: bird, high, eye, low, worm *yaw*: front, profile, back
on-screen position	*horizontal*: left, center, right *vertical*: upper, middle, lower
body part visibility	*partial*: legs, foot, hand *hidden*: head, eyes, torso, legs
artistic style	*framing*: cowboy, OTS, OTH, FS, choker, dutch *actions*: lying, sitting, acrobatic, dummy

4. Evaluation framework of HPE models in connection with film character representation

In this section, we describe the methodology to assess the performance of existing HPE models on film content. We first motivate our choice of specific HPE models in Section 4.1. We then present in Section 4.2 the datasets we consider to study HPE in connection with character representation, and motivate the introduction of a new dataset we name TRACTIVE. Notably, as no existing dataset available for HPE assessment has been yet annotated with cinematographic style, we describe our annotation effort, consisting in a frame sampling process and two pieces of software.

4.1. Selected HPE methods

We choose to benchmark on film data two HPE methods presented in Section 2.3. We select OpenPose [CSWS17] because it is a reference bottom-up approach shown to have reliable and real-time performance, and DEKR [GSX*21] because (i) it is a most recent approach shown to outperform existing competitors and because (ii) it builds on the features learned by a top-down approach, HRNet, to take the best of both bottom-up and top-down approaches.

4.2. Selection and creation of datasets

We consider two existing datasets and a new dataset as described below.

COCO2017 [LMB*14] Composed of 106K images that are annotated with 17 joints, as introduced in Section 2.3. Both DEKR and OpenPose are trained on the COCO2017 training dataset, since it is the only dataset with full-body pose annotations. The ground truth (GT) labels for the test dataset are not made available, so we use the validation dataset instead, composed of 5K images, to conduct the performance analysis on COCO2017 in Section 5.4.

FLIC [ST13] We consider the FLIC-plus dataset made of 17K images, as introduced in Section 2.3. In contrast withh COCO2017, this dataset is entirely made of professional film shots, but is annotated with only 10 upper-body joints and ground-truth poses for up to two characters only (owing to the human annotation being conducted only on characters first detected using the Poselets person detector [BM09]). Both of these factors will result in characters in the image being systematically left out of the annotation, as well as smaller bounding box sizes. Due to these limitations, we were not able to train the models on FLIC.

TRACTIVE As our higher-level motivation is to study varied character representation in films, notably corresponding to the concepts of gender, *male gaze* and *female gaze*, we introduce a new dataset of clips extracted from the corpus proposed by Brey [Bre20]. This dataset is composed of 13 clips, on which HPE methods are evaluated on a subset of frames hand-annotated for both character representation and pose estimation quality, as detailed below.

4.3. Annotation process

Two requirements need to be fulfilled for the annotation process:

[R1] Estimation of the quality of HPE: If the ground truth of character bounding boxes is available in the original dataset, then nothing needs annotation. Else, hand annotation of the HPE quality is required.

[R2] Annotation of character representation: To be done for all three datasets.

(a) *OTS: over the shoulder* **(b)** *OTH: over the hip* **(c)** *FS: full shot head to toe* **(d)** *Choker: face fills screen* **(e)** *Dutch: rolled camera*

Figure 1: *Examples of artistic shots from Table 1.*

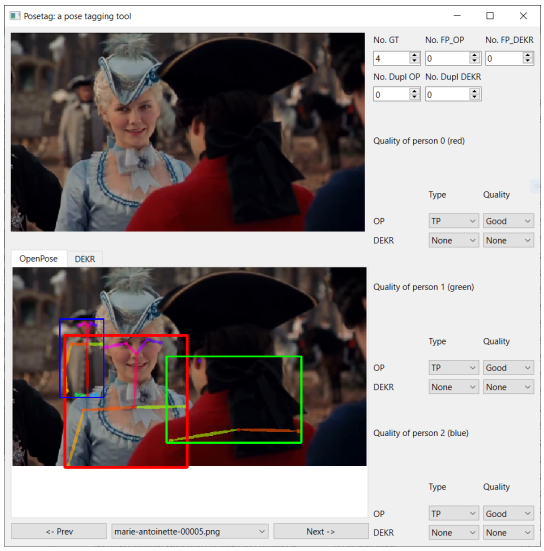

Figure 2: *The Posetag annotation tool that allows the user to assess pose detection results by indicating if detected poses are true or false positive, and a quality score (good, medium, bad).*

To make the annotation task feasible and efficient, we first sample a subset of frames with a wide variety in style, dataset, and challenge to pose detection algorithms. We targeted a subset of 1000 images each from the COCO2017 and FLIC datasets. For each frame, we calculate:

- N_GT: the number of human-annotated ground truth poses,
- N_TP: the number of true positive detections by the HPE based on Intersection Over Union (IOU) metric – a threshold of the ratio between detected pose bounding box and GT bounding box,
- Precision and Recall: on character detection, calculated as $N\_TP/(N\_TP+N\_FP)$ and N_TP/N_GT respectively, with $N\_FP = N\_GT - N\_TP$
- Bounding box (bbox) size: we take the ratio of the largest ground truth bounding box to the image, and the ratio between the largest pose-detected bbox to the largest GT bbox.

For COCO, we sample 60 images each for the maximum, minimum, and a range of values between min and max for each criterion. For FLIC, we sample 2 images per min-max-range criterion for each film. The unique set of all the sampled images resulted in 1018 images out of the original 5000 images for COCO2017 and 1244 images out of the original 17K for FLIC. For TRACTIVE, we sampled 500 images from the 13 clips at fixed time intervals, removing close duplicates.

The selected subsets of images from all three datasets were then annotated by a human expert (one of the authors) with the relevant character representation labels from Table 1, using two pieces of software we developed for this purpose: [R1] is addressed by the Posetag tool shown in Figure 2, to annotate the quality of automatic pose estimation (whether it is ground truth, and quality score), used on the TRACTIVE dataset where no pose ground truth is available, and [R2] is addressed by the Cinetag tool, which has instead a right panel of checkboxes to select the character representation labels from Table 1 for each image, annotated for all three datasets.

We thereby obtain a total of over 2700 images with pose detection results from two HPE methods and labels for character representation, serving as the basis of our subsequent analyses.

5. Experimental results and analysis

The results of OpenPose and DEKR on the three datasets has resulted in a rich set of metrics for analysis. We present the key findings in this section, starting with the general characteristics of the two pose detection algorithms on all three datasets, and then focusing on the results of OpenPose to discuss the impact of cinematographic features. Finally, we present a few salient examples of difficult cases.

5.1. Training configuration and evaluation metrics

Owing to the lack of keypoint annotation in TRACTIVE, and the poor quality of keypoint annotation in FLIC (only upper-body parts, at most two characters per image with numerous unlabeled characters and redundant labels), we consider both methods Openpose and DEKR trained on the COCO2017 dataset. From the respective websites making these methods available, we re-train the provided models and verify that similar performance are obtained on the COCO2017 test set (whose labels are not disclosed publicly).

To analyze the results of the HPE methods on various datasets, we resort to:

- the criterion N_GT excluding images that have $N\_GT = 0$,
- the criteria of cinematographic features described in Table 1
- the metrics of Precision and Recall defined above. They are obtained from GT annotation: from the original dataset in COCO2017 and FLIC, and from our annotation process described above for TRACTIVE.

5.2. Quantitative average results

At a high level, we have two goals: (1) compare quantitative performance of the two HPE methods on each dataset, and (2) compare the performance between the three datasets, and analyze their respective composition in terms of character representation.

5.2.1. Comparison of HPE methods: OpenPose and DEKR

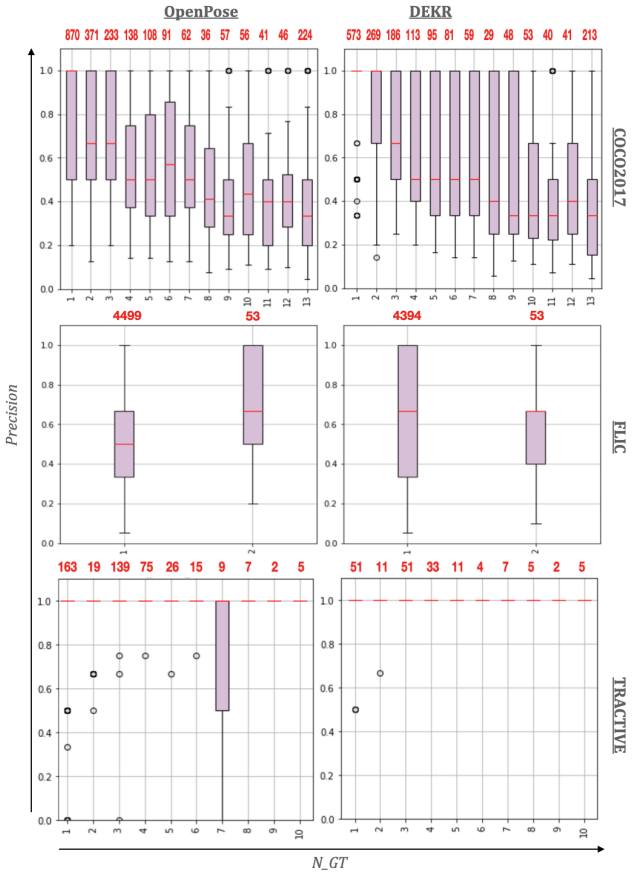

Figure 3: *Boxplots of precision values (wrong detections) for a group of images, the images being grouped by the number of visible characters N_GT . Comparison of the precision of OpenPose and DEKR on each of the three datasets. The median is indicated by the red line. The number of images for each boxplot is in red text, which excludes those where no poses are detected. Wrong detections are more frequent when the number of characters increases in COCO2017, but remains stable in the other two datasets.*

Overall, as shown in Figure 3, precision vs. N_GT show equivalent results for both HPE methods, with a decrease in precision as N_GT increases. The number of samples reflects the amount of images where at least one pose was detected, which is much higher in OpenPose. When we look at recall in Figure 4, DEKR performs systematically worse than OpenPose, for the same detection threshold of 0.8 for both. Decreasing the threshold improves recall of DEKR at the expense of lower precision. DEKR also frequently has duplicate detections for the same character depending on the IOU threshold. For both HPE methods, scatter plots (not shown here for space limitation) show a plateau in recall for $N\_TP \geq 5$ characters, above which the number of detected characters does not really increase.

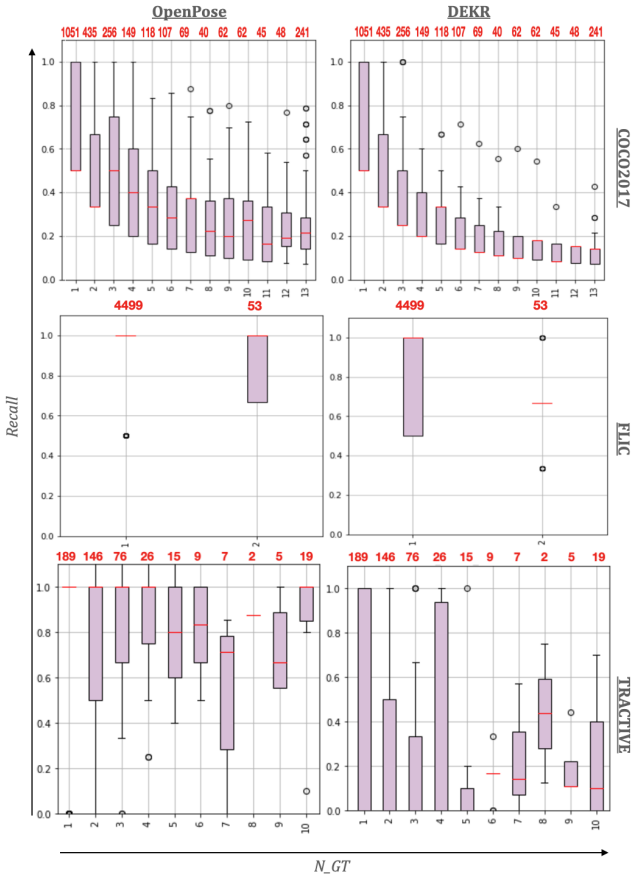

Figure 4: *Boxplots of recall (missed detections) values, the images are grouped by the number of visible characters N_GT. Comparison of the recall of OpenPose and DEKR on the three datasets. Figure shows that missed detections generally increase with the number of characters on screen. OpenPose also outperforms DEKR on the film datasets FLIC and TRACTIVE.*

5.2.2. Comparison of performance between the three datasets

On all three datasets, we generally observe in Figure 3 and 4 that precision and, even more obviously, recall decrease with N_GT, indicating proportionally more misses with the number of on-screen characters. Figure 4 shows that in more than 25% of the cases, at least 20% of characters are missed for $N\_GT \geq 2$. For $N\_GT \geq 5$, at least 50% of characters are missed. COCO is the most comprehensively annotated dataset amongst the three, and the lowest recall is obtained on COCO. However, like for many hand annotated datasets, in COCO there are a number of false annotations [XBG*19] and noise in the pose data [NNV19], such as humans labelled when none are visible, which negatively impacts recall.

The number of annotated N_GT in FLIC is lower than the actual number of on-screen characters, limited to either 1 or 2 characters,

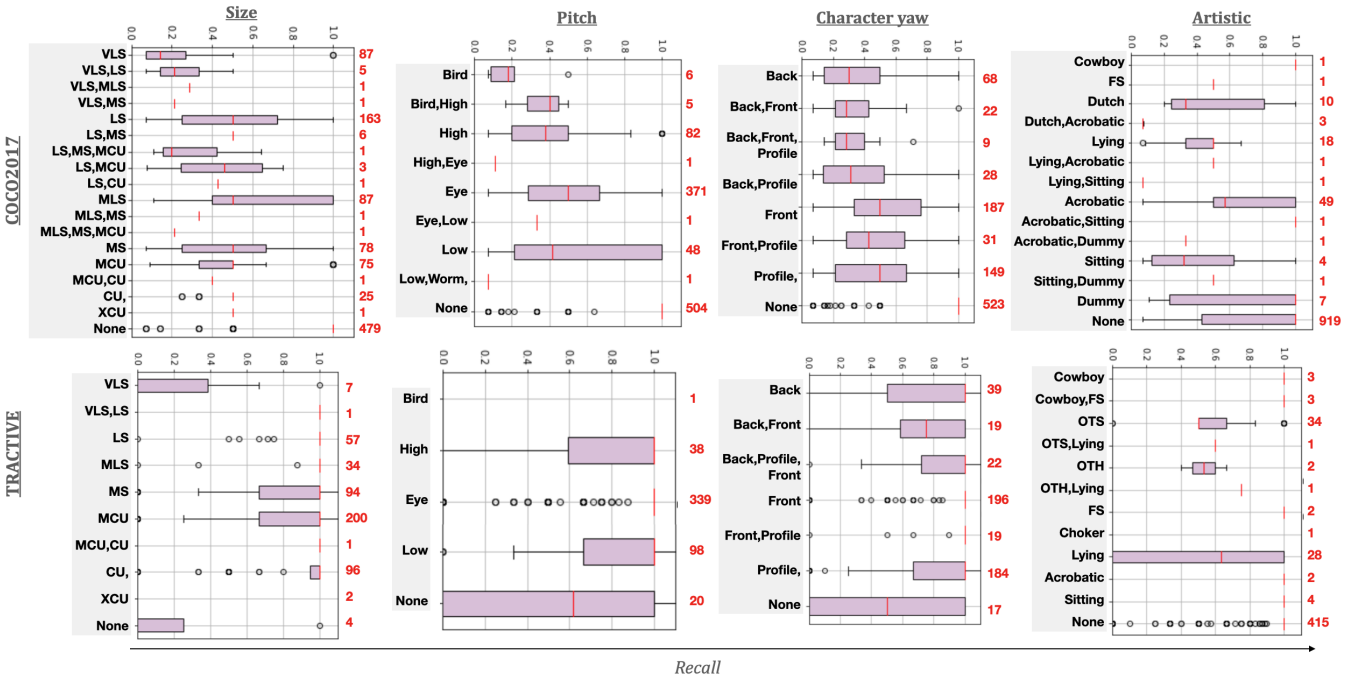

Figure 5: *Boxplots of recall (x-axis) values on OpenPose for each image, grouped by cinematographic features of size, angle (pitch and yaw) and artistic tag. The number of images for each boxplot is in red text. Multiple tags (y-axis) represent more than one main character on screen. We see that missed detections are most frequent with extreme shot sizes, non-horizontal pitch, non-front yaw, and non-upright poses.*

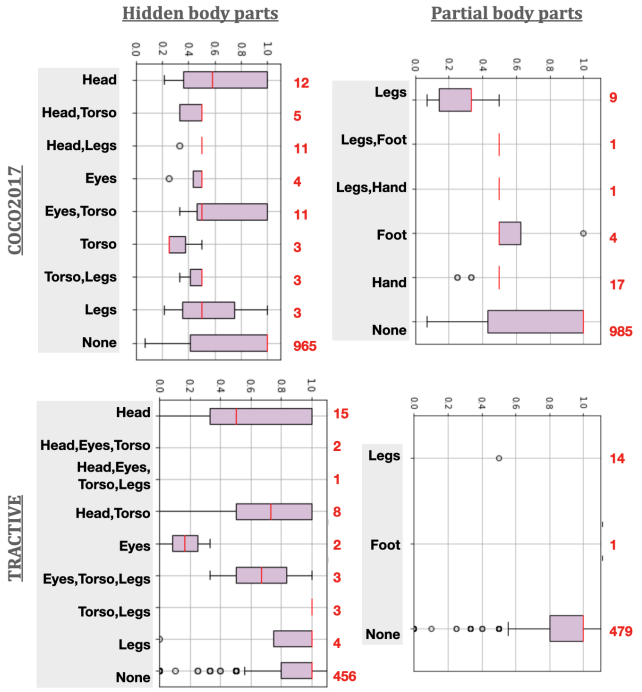

Figure 6: *Boxplot of recall per image, images grouped by hidden (left) and visible (right) body parts. Figure shows that missed detections are almost systematic when only headless body parts are visible (right column).*

as shown in Figures 3 and 4. Detections by HPE corresponding to actual characters are hence often tagged as FP, artificially lowering precision, and raising the recall. Additionally, ground truth annotations only have the upper body, which can result in correct detections being labelled as FP due to IOU mismatch. We can observe a globally higher precision on TRACTIVE than other datasets, which is the result of fewer false positives compared to COCO and FLIC. This is both because TRACTIVE has a more precise annotation of number of characters, and no IOU criterion to verify the correspondences of the detected and ground truth bounding boxes. On TRACTIVE, when N_GT<5 all characters are correctly detected in more than 50% of the frames.

5.3. Impact of film style on results

We now break down the results by the cinematographic features concerning character representation, and analyze their impact on HPE performance. We conduct this analysis on OpenPose, which has a more stable performance in multi-character frames, and on the COCO and TRACTIVE datasets, due to over-estimation of recall in FLIC. The collective results of size, angle (pitch and yaw), and number of characters are shown in Figure 5 and the comparison of recall to features pertaining to body part visibility are in Figure 6.

Effect of size. Recall drops for VLS, as expected since the characters are less identifiable. Even more interesting is the lowering of recall for closer shots, too: MS, MCU, CU and XCU, which is a key finding that questions the efficacy of such methods for film

shots. It is worth noting that these shots are the least frequent in COCO on which the methods have been trained.

Effect of camera angle. Recall decreases in both COCO and TRACTIVE whenever the pitch is not eye level, or when the character yaw is not towards the front (i.e., profile or back view).

Effect of artistic framing. Lying characters are significantly more difficult to detect, with a non-negligible number of occurrences. We can see in TRACTIVE that a major difficulty in artistic shots arises with OTS and OTH, where the closer character is almost systematically missed. This is another key finding where the methods must be improved to work properly for film content. Similarly to shorter shot sizes, neither OTS nor OTH may be present in the training set, as they are not in the sampled annotation set from Section 4.3.

Effect of number of characters. In coherence with *N_GT*, recall decreases with the increase in *nChar*. In our annotations, there can be two labels for *nChar*: the first for foreground and the second for background characters.

Effect of body part visibility. Any hidden body parts result in lower recall, for both TRACTIVE and COCO, though there is still some resiliency when the head and/or eyes are hidden. There are not many cases where the head is not visible (23 in COCO). It is also important to note that the detected characters in these images are not necessarily the ones with hidden body parts. When only a single or lower-body parts – usually legs, feet or hands – are visible, the recall is zero. There are 14 occurrences of images with only legs in TRACTIVE. It is therefore key to incorporate proper body part detection into an HPE framework for film shots, which can be significant when we consider gender representation and how the body is portrayed in films.

Effect of on-screen position. While HPE methods, and more generally object detection models, are robust in the position of targets in the image, we find that characters appearing in the lower part of the image can often be cut off by the frame edge, and characters on the left and right side of the frame would more often be in profile position instead of facing the camera, resulting in lower precision / recall that is indirectly linked to on-screen position.

While we analyze each category of features separately, combinations of features could provide more refined information on HPE weaknesses. Also worth noting is that artistic shots often have multiple features that impact precision and recall. For example, OTS shots are usually close shots with a back character yaw. Cowboy shots systematically cut characters off at the knees.

5.4. Qualitative analysis of failure cases

Here we collect a number of examples outlining difficult cases for HPE methods. We can see from Figure 7 a selection of these examples. Firstly, short shots often result in poses with missing points as in (a) for the shoulders, or completely missing as in (i). The opposite, very long shots also pose a difficulty as in (b) and complex lighting or architecture such as pillars can also result in false positives as in (e). The same goes for over-the-shoulder (OTS) shots where characters are only partially detected in (c) or not at all in (d). Importantly, many shots where characters are objectified may often only show lower-body parts in the entire frame, as in (f) and (h), and are not recognized as poses by the algorithms. This is in contrast to occlusion, which is a better addressed question, and poses less of a problem for modern HPE methods, as seen in (j).

6. Discussion

From the preceding analysis, we can see both strong potentials and limitations of deep pose detectors for qualitative film analysis. In this post-analysis discussion, we would like to address two main points: the potential of HPE-assisted annotations for film datasets and the role of cinematographic style in training deep pose detectors. Finally, we discuss the significant role HPE methods can play in quantitative analysis of character representation.

First of all, we have found that while pose detection has its limitations, it can greatly reduce the annotation load of human experts. Out of the 500 images in the TRACTIVE dataset, around 423 images had at least one pose detection by OpenPose which was deemed as medium quality or above, around 300 of which deemed "Good" (no missing keypoints). The value of pose detection-assisted annotations is clearly exploited in the creation of the FLIC dataset, and with more robust detectors built on deep architectures, we can imagine the annotation load to be significantly reduced, and accuracy increased.

Second, to our knowledge, this is the first paper to identify the impact of cinematographic style on HPE methods. From the analysis in Section 5.4 of the expert annotations, we can conclude that elements of character representation including size, angle, occlusion and frame border cutoffs, and artistic framings – in particular, OTS, OTH, and partial body poses – all present difficult cases for HPE methods, but are extremely common in film content. We also observed that current existing datasets do not provide sufficient training data for these types of shots. Therefore, new datasets must be established, and/or the existing ones expanded to train deep learning algorithms that can carry out artistic interpretations or classification for film data.

7. Conclusion

In this paper, we have evaluated the performance of the deep pose detectors DEKR and OpenPose from the perspective of film style and character representation. In this workflow, we have also identified the added value of the cinematographic style labels of character representation, which currently could only have been obtained through human annotations. The advantages and limitations offered by HPE methods are twofold. On the one hand, they could be a valuable tool in the automatic and quantitative analysis of character representation in film. However, on the other hand, to achieve this goal, the training process must include images and data with a wider range of character representations, particularly OTS and OTH shots, and shots with only lower-body parts. We envision two main directions to improve the performance of HPE methods on film data. The first direction is to augment the training data with problematic shots. Thanks to the present work, the annotation effort can be focused on specific types of shots to strengthen the training data using our developed tools. While the most straight-forward approach, this still involves costly human annotation. An alternative, or complement, will be to crop existing annotated shots to train the

(a) *TRACTIVE - OTS* **(b)** *TRACTIVE - Bird view* **(c)** *TRACTIVE - legs only* **(d)** *TRACTIVE - lying, legs* **(e)** *TRACTIVE - XCU*

Figure 7: *Examples of difficult cases in the dataset for pose detection. (a) features an OTS shot of a character with their back towards the camera; (b) is a bird eye angle with false detection of pillars as humans; (c)(d) show the weakness for shots with only partial body parts; (e) is an extreme closeup shot.*

models to recognize poses with missing upper-body parts, which we have identified as a major difficulty. Also, the annotation effort may be mitigated with active learning strategies, for example based on out-of-distribution sample detection [MKvA*21]. Finally, methods like OpenPose rely on individual body part detection and assembly. It will therefore be interesting to derive new model architectures better suited at pose estimation from individual body part detection.

Acknowledgements

This work has been supported by the French National Research Agency through the ANR TRACTIVE project ANR-21-CE38-00012-01, the EUR DS4H Investments in the Future projects ANR-17-EURE-0004, and by EU Horizon 2020 project AI4Media, under contract no. 951911 (https://ai4media.eu/).

References

[APGS14] ANDRILUKA M., PISHCHULIN L., GEHLER P., SCHIELE B.: 2d human pose estimation: New benchmark and state of the art analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014), pp. 3686–3693. 2

[BLC*21] BAO Q., LIU W., CHENG Y., ZHOU B., MEI T.: Pose-Guided Tracking-by-Detection: Robust Multi-Person Pose Tracking. *IEEE Transactions on Multimedia 23* (2021), 161–175. 2

[BM09] BOURDEV L., MALIK J.: Poselets: Body part detectors trained using 3d human pose annotations. In *2009 IEEE 12th International Conference on Computer Vision* (2009), pp. 1365–1372. 3

[Bre20] BREY I.: *Le regard féminin : Une révolution à l'écran*. Points, 2020. 3

[CLCK21] COURANT R., LINO C., CHRISTIE M., KALOGEITON V.: High-level features for movie style understanding. In *ICCV 2021 Workshop on AI for Creative Video Editing and Understanding* (2021). 2

[CSWS17] CAO Z., SIMON T., WEI S.-E., SHEIKH Y.: Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI, July 2017), IEEE, pp. 1302–1310. 2, 3

[GSX*21] GENG Z., SUN K., XIAO B., ZHANG Z., WANG J.: Bottom-Up Human Pose Estimation Via Disentangled Keypoint Regression. In *2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Nashville, TN, USA, June 2021), IEEE, pp. 14671–14681. 2, 3

[HXR*20] HUANG Q., XIONG Y., RAO A., WANG J., LIN D.: MovieNet: A Holistic Dataset for Movie Understanding. In *European Conference on Computer Vision (ECCV)*, Vedaldi A., Bischof H., Brox T., Frahm J.-M., (Eds.), vol. 12349. Springer International Publishing, Cham, 2020, pp. 709–727. 3

[LMB*14] LIN T.-Y., MAIRE M., BELONGIE S., HAYS J., PERONA P., RAMANAN D., DOLLÁR P., ZITNICK C. L.: Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)* (2014), Springer International Publishing, pp. 740–755. 2, 3

[Mas65] MASCELLI J. V.: *The five C's of cinematography*, vol. 1. Grafic Publications, 1965. 2

[MKvA*21] MUKHOTI J., KIRSCH A., VAN AMERSFOORT J., TORR P. H. S., GAL Y.: Deterministic Neural Networks with Inductive Biases Capture Epistemic and Aleatoric Uncertainty. *arXiv:2102.11582 [cs, stat]* (June 2021). arXiv: 2102.11582. URL: http://arxiv.org/abs/2102.11582. 8

[NNV19] NEVEROVA N., NOVOTNY D., VEDALDI A.: Correlated uncertainty for learning dense correspondences from noisy labels. *Advances in Neural Information Processing Systems 32* (2019). 5

[RVB13] RONFARD R., VINEET G., BOIRON L.: The prose storyboard language. In *AAAI Workshop on Intelligent Cinematography and Editing* (2013), vol. 3, Citeseer. 2

[SGM*21] SOMANDEPALLI K., GUHA T., MARTINEZ V. R., KUMAR N., ADAM H., NARAYANAN S.: Computational media intelligence: Human-centered machine analysis of media. *Proceedings of the IEEE 109*, 5 (2021), 891–910. 2

[ST13] SAPP B., TASKAR B.: MODEC: Multimodal Decomposable Models for Human Pose Estimation. In *2013 IEEE Conference on Computer Vision and Pattern Recognition* (Portland, OR, USA, June 2013), IEEE, pp. 3674–3681. 2, 3

[SXLW19] SUN K., XIAO B., LIU D., WANG J.: Deep high-resolution representation learning for human pose estimation. In *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 5686–5696. 2

[WGLC17] WU H.-Y., GALVANE Q., LINO C., CHRISTIE M.: Analyzing elements of style in annotated film clips. In *Eurographics Workshop on Intelligent Cinematography and Editing (WICED)* (2017), The Eurographics Association, pp. 29–35. 2

[WPRC18] WU H.-Y., PALÙ F., RANON R., CHRISTIE M.: Thinking like a director: Film editing patterns for virtual cinematographic storytelling. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 14*, 4 (2018), 1–22. 2, 3

[XBG*19] XU M., BAI Y., GHANEM B., LIU B., GAO Y., GUO N., YE X., WAN F., YOU H., FAN D., ET AL.: Missing labels in object detection. In *CVPR Workshops* (2019), vol. 3. 5

[Zet16] ZETTL H.: *Sight, sound, motion: Applied media aesthetics*. Cengage Learning, 2016. 2

[ZWCS20] ZHANG Y., WANG Y., CAMPS O., SZNAIER M.: Key Frame Proposal Network for Efficient Pose Estimation in Videos. In *European Conference on Computer Vision (ECCV)* (Cham, 2020), Vedaldi A., Bischof H., Brox T., Frahm J.-M., (Eds.), vol. 12362, Springer International Publishing, pp. 609–625. 2

[ZWY*20] ZHENG C., WU W., YANG T., ZHU S., CHEN C., LIU R., SHEN J., KEHTARNAVAZ N., SHAH M.: Deep learning-based human pose estimation: A survey. *arXiv preprint arXiv:2012.13392* (2020). 2