Zero-shot neuro-symbolic parsing of body keypoints

Dalia Aljawaheri, Gertjan Burghouts, Jan Erik van Woerden, Judith Dijk, Hugo Kuijf
TNO, the Hague, the Netherlands

ABSTRACT

A new approach for distinguishing neutral (e.g. walking) from threatening (e.g. aiming a handgun) poses, without training, is presented. There are various AI-based models that can classify human poses, but these oftentimes do not generalize to defence scenarios. The lack of data with threatening poses makes it hard to train new models. Our approach circumvents re-training and is a zero-shot, rule-based classification method for threatening poses.

We combine a pretrained body part keypoint detection model with the neuro-symbolic framework Scallop. We compare the pretrained models MMPose and YOLOv8x-pose for keypoint detection. We use images from the YouTube Gun Detection Dataset containing persons holding a weapon and label them manually as having a 'neutral' or 'aiming' pose; the latter was further subdivided into 'aiming a handgun' and 'aiming a rifle'. Scallop is used to define logic-based rules for classification, using the keypoints as input: e.g. the rule 'aiming a handgun' includes 'hands at shoulder height' and 'hands far away from the body'.

Recall and precision results for aiming are 0.75/0.81 and 0.83/0.73, for MMPose and YOLOv8x-pose, respectively. Average recall and average precision for 'aiming a handgun' and 'aiming a rifle' are 0.78/0.36 and 0.76/0.43, for MMPose and YOLOv8x-pose, respectively.

Combining neuro-symbolic AI with pretrained pose estimation techniques shows promising results for detecting threatening human poses. Performance of neutral-versus-aiming classification is similar for both approaches, however, MMPose performs better for multi-class classification. In future research, we will focus on improving rules, identifying more poses, and using videos to obtain sequences of poses or activities.

Keywords: machine learning, artificial intelligence, human pose estimation, neuro-symbolic, fine-grained classification, EDA DEBELA

1. INTRODUCTION

Detecting a ground-based threat before it materializes is critical, so that defensive actions can be taken in a timely manner [1]. For example, when someone is aiming a weapon like a gun or a rifle. Artificial intelligence (AI)-based computer vision technology can assist with this task, by employing human pose estimation techniques that can recognize a threatening pose in images or videos. For various computer vision tasks, state-of-the-art AI techniques already achieve performance near, or even better than, humans.

Despite having such high performance, commonly available human pose estimation (or action recognition) techniques are trained for day-to-day activities, such as walking, running, cycling, etc. Moreover, they often assume that there is a clear and close-by view on the subject of interest, which is often not the case in defence and security applications. Another frequent problem is the generalization of AI-based techniques that do not easily translate to new tasks without re-training on a new and large dataset. Such datasets are not readily available for human poses associated with threatening actions, such as aiming a weapon.

Several deep learning-based methods have attempted to detect aiming persons and weapons. Detecting just a weapon is generally a difficult task, since the weapon itself is a minor fraction of the total image size [2]. Several methods have aimed to overcome this problem by trying to improve the detector [3], however this is still the largest source of inadequate performance for weapon detection [4]. Occlusion (partial/complete visibility) of the weapon would even further complicate detection in such cases [5]. Additionally, detecting different weapons than the weapons the model is trained with, is another challenge and can lead to a considerable number of false positives [6].

A new development is the use of additional information, such as pose estimation, that can be used for weapon classification. A recent approach [7] tries to optimize the gun detection based on the hand gesture. They train a hand landmarks detection

Electro-Optical and Infrared Systems: Technology and Applications XXI, edited by
Duncan L. Hickman, Helge Bürsing, Ove Steinvall, Philip J. Soan, Proc. of SPIE
Vol. 13200, 132001C · © 2024 SPIE · 0277-786X · doi: 10.1117/12.3031487

model, which is followed by a customized feature selection method and a CNN as classifier; classifying whether the hand is holding a gun or a non-threatening object. The authors report a 93% accuracy using their own Face-and-Gun-Dataset, in which the hands and objects are the central part of the image; they do not deal with difficulty of having the weapon as only a small object in the image.

Other models have therefore tried using human pose as additional information. One approach [6] used the OpenPose model, which extracts body keypoints. Based on these keypoints, small image patches of the hands are obtained. The image patches are processed with a CNN to obtain visual features, while the keypoints are processed by a separate feed-forward network to obtain pose features. The two obtained collections of features are combined, and a final classifier determines if a handgun is present or not. This model needs training twice, because the feed-forward network is first trained to classify between 'dangerous' and 'other' poses, while the final classifier needs to be trained to classify 'handgun' or 'no handgun'. They report to achieve an average precision of 85-92% (depending on the test dataset).

Another similar approach, using human pose and hand region information, is proposed by [8]. The authors use a body keypoint detection model that, amongst others, outputs keypoints for the elbows and wrists. Based on these keypoints, they calculate where the hand region is and keep only the two crops of the hands. Then a weapon detection model, trained only on weapons, detects the weapon. Depending on the architecture of the weapon detection model, they report a precision of 91%.

Although these methods, and other similar ones, show promising results, there still is a general difficulty with the small number of pixels on a weapon when the person is not the main object of the image. All methods need training and therefore have a smaller flexibility regarding other weapons. A zero-shot solution that avoids re-training and uses the human pose could address such limitations and provide a solution.

In this work, we propose such a zero-shot approach to classify threatening human poses (aiming a weapon) versus non-threatening human poses. First, we compare two state-of-the-art human body keypoint detection models that can detect body parts in images, namely MMPose and YOLOv8x-pose, to compare their impact on the final classification performance. The neuro-symbolic framework Scallop is used to incorporate logic-based rules about threatening poses, which are based on the found body keypoints. With the logic-based rules we incorporate expert domain knowledge into our approach and thereby avoid training on the specific poses. With this probabilistic rule-based classification approach, we can distinguish threatening from non-threatening poses and further classify the threatening pose into two distinct classes, namely 'aiming a handgun' and 'aiming a rifle'.

2. METHODS AND MATERIALS

Our zero-shot approach combines a body part keypoint detection model with the neuro-symbolic framework Scallop for classification of the human pose. For body keypoint detection we use two pretrained models, MMPose and YOLOv8x-pose, which return 17 body keypoints. We use these two models to compare the impact of their performance on the final classification. In Scallop, we define logic-based rules that describe the threatening poses. Using the probabilistic, detected body keypoints as input for the rules, Scallop can classify the pose as threatening or not. Figure 1 shows the overall approach.

Datasets

To evaluate our method, we used two datasets: (1) the YouTube Gun Detection Dataset (YouTube-GDD) [9] and (2) a dataset recorded at a field-trial of the European Defence Agency (EDA) CAT B project, Detect Before Launch (DEBELA) [1]. The YouTube-GDD will be used for quantitative evaluation of the proposed method and the DEBELA dataset for a qualitative evaluation.

The DEBELA dataset was recorded at WTD-52, Oberjettenberg, Germany, as part of a field trial that was organized in June 2023 [1]. One of the aims of the DEBELA-project is to evaluate image analysis technology to detect ground-based threats before they launch, for example people aiming a weapon. This dataset contains, amongst others, three different scenarios of people aiming a weapon in the direction of the camera. Each scenario has a different background and field-of-view. The three videos that were extracted from this dataset will be used for a qualitative evaluation of our approach and to estimate its applicability on domain-specific data.

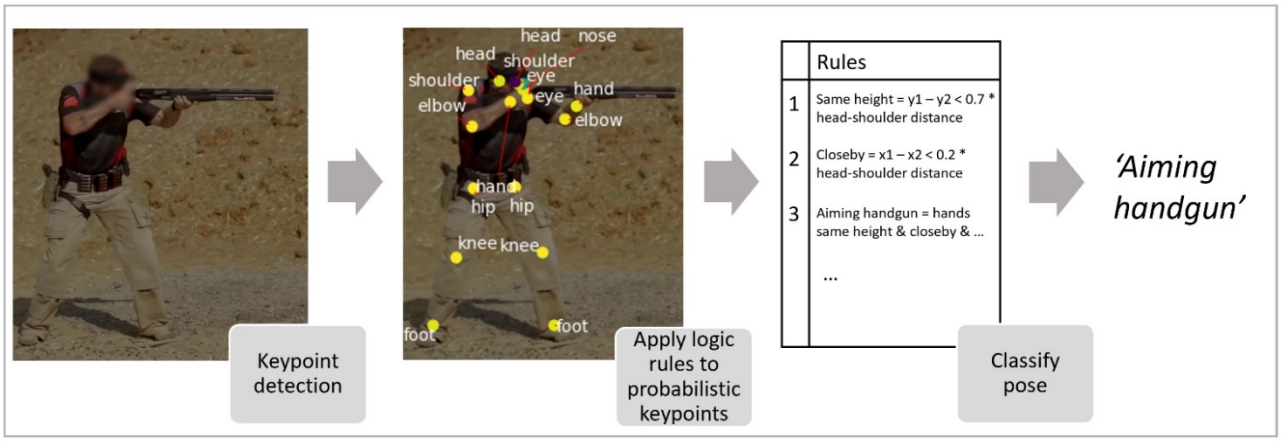

Figure 1. Our approach for threatening pose classification. With a keypoint detection model, MMPose or YOLOv8x-pose, 17 keypoints are detected. Using the neurosymbolic framework Scallop, logic rules are applied to these keypoints to identify a threatening pose.

Table 1. The 17 keypoints as defined in the COCO-dataset and used by YOLOv8x-pose and MMPose. L = left. R = right.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Nose	Eye (L)	Eye (R)	Ear (L)	Ear (R)	Shoulder (L)	Shoulder (R)	Elbow (L)	Elbow (R)
10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	
Wrist (L)	Wrist (R)	Hip (L)	Hip (R)	Knee (L)	Knee (R)	Ankle (L)	Ankle (R)	

Keypoint detection

Methods for 2D human body keypoint detection perform two tasks at the same time: detection of a person and localization of important keypoints, such as the head, hands, knees, etc. Detection and localization of these keypoints in an image can be helpful when estimating someone's pose and possibly the action associated with it. Commonly available datasets with annotated human body keypoints differ in the amount of keypoints provided, ranging from very specific locations (e.g. the face) to full body. The Common Objects in Context (COCO [10]) dataset is a large and well-known dataset that covers over 200,000 images with 17 keypoints on the body (see Table 1). This dataset is used by several state-of-the-art methods to create pre-trained human body keypoint detection methods that can be applied in a zero-shot setting (i.e. without retraining on a specific dataset). Two of such methods will be used and compared in this work: YOLOv8x-pose [11] and MMPose's HRNet [12]. Both models are well-performing, pretrained, and easily available online.

YOLOv8x-pose

The YOLO models are widely known for their good zero-shot performance on various object detection tasks [11]. Recently, YOLOv8 was made available, which can perform multiple tasks including human pose estimation by extracting body keypoints [13]. The model is trained on the COCO keypoints dataset and is available in five model sizes. We use the largest model, YOLOv8x-pose, since this model shows the best performance as reported in literature. The model outputs the x- and y- coordinates of the 17 detected keypoints and the corresponding probability. The original authors report a mean average precision (mAP) of 90.2 at 50% Intersection over Union (IoU) on the COCO val2017 dataset [14].

MMPose

MMPose is an open-source toolbox, part of the OpenMMLab project. The toolbox contains, amongst others, a set of pre-trained models for human pose estimation. We used the state-of-the-art architecture HRNet [15], specifically HRNet-W48, which aims to have a higher accuracy than its smaller counterparts (e.g. W32) and is pre-trained on the same COCO dataset. Similarly, it outputs the same 17 keypoint x- and y- coordinates for the human body with the corresponding probability. The model is reported to have a mAP of 90.6 on the COCO val2017 dataset.

Figure 2. Two examples from the YouTube-GGD. Left: an image labelled as 'aiming a handgun'. Right: an image labelled as 'aiming a rifle'.

Neuro-symbolic framework

Scallop is a neuro-symbolic framework [16] that can combine logic rules with probabilistic statements. Commonly, Boolean logic rules are a structured way of defining factual knowledge using three logical operators: 'AND', 'OR' and 'NOT'. Boolean logic can only be 'True' or 'False' and thus is binary. However, the output of deep learning models is oftentimes probabilistic. It is desirable to combine the output of deep learning networks, such as keypoint detections with a probability, to predefined domain-specific knowledge rules. To combine these two worlds, Scallop propagates the probabilities through logic rules using so-called provenance semirings. In our case, we use the min/max probability provenance. Of key importance is that when probabilities are combined in a logic rule by the operator 'AND', the minimum of the probabilities is taken; while for handling the operator 'OR' the probabilities are combined through the maximum. This is especially useful for scenarios when there are multiple possibilities, and the possibility with the highest probability is preferred.

Classification logic rules

The Boolean logic rules in Scallop are the part where domain-specific knowledge can be incorporated. In our case, we aim to identify people aiming a weapon, where initially it is not important to classify the type of weapon. We also do not wish to find these people based on a weapon detection method—since this is not trivial owing to the large variety in weapons and the usually limited number of pixels-on-target—but predict the threatening intent by recognizing their body positioning when aiming a weapon. To cover the variety of aiming positions (because of different types of weapons), we focus to cover the two aiming positions shown in Figure 2.

The rules to describe these two poses are based only on the location of the keypoints for the hands, elbows, and shoulders. Since images can have a narrow or wide field-of-view, relative distances within the image are obtained by normalizing with the distance between the keypoints for the ears and shoulders.

To define accurate body positions, a reference measurement is needed to calculate relative keypoint distances. Since this cannot be measured in 'number of pixels'—because of the non-constant image scales and perspectives (camera closer or farther from the target person) and hence the number of pixels on the person varies—we obtain relative distances by normalization. To do so, we use the maximum distance between the ears and shoulders as a normalization factor; we take the largest distance between two y-coordinates (one ear and one shoulder coordinate) and define this as our ES (ear-shoulder) distance, which we will use in the following two body position definitions.

Below, the descriptions of the two rules defined in Scallop are provided. We implemented the rules in Python, since Scallop provides a Python interface that allows adding rules and combining these with the probabilistic deep learning outcomes. The combination of these two rules provides the classification 'aiming a weapon'.

- *Body position 1 ('aiming a handgun')* When aiming a handgun, generally the arms are stretched out before the body and both hands hold the gun. Therefore, we define 'aiming a handgun' when the two hands, elbows, and shoulders are at the same height and both hands are close to each other. Having the same height here is defined as 0.7 × ES distance. The factor 0.7 is found heuristically to achieve optimal results. Being close by is defined as 0.2 x ES distance, where the factor 0.2 is also found heuristically.

- *Body position 2 ('aiming a rifle')* For this position, we again state that the hands and shoulders are at similar height (relatively, using the ES distance), however the two elbows are not. Since there is a variety in rifles, where hands can be either close or far away, we put no restrictions on the distance of the two hands.

3. RESULTS

This section describes our experiments and results. First, both YOLOv8x-pose and MMPose are used to extract keypoints from the images of the YouTube-GDD, after which our rule-based classifier in Scallop classifies all images into 'aiming' versus 'not aiming'. The two keypoint detection models are compared to evaluate which model, combined with Scallop, results in the best classification performance. The classification performance is evaluated using recall and precision. To mimic the harder setting of defence and identify the limits of our classification approach, the image resolution of the YouTube-GDD will be reduced. Based on these results (MMPose outperforms YOLOv8x-pose), we process the videos of the DEBELA-dataset using MMPose with Scallop and provide qualitative results.

YouTube-GDD – Classification – full resolution

We use recall and precision to evaluate the classification performance of our method, comparing the impact of each keypoint detection model (YOLOv8x-pose and MMPose) separately, when combined with the rule-based classification in Scallop. Results for the classification of 'aiming' versus 'not aiming' are provided in Table 2; together with the classification results for 'aiming a handgun' and for 'aiming a rifle'.

Both keypoint detection methods show similar results for the 'aiming' classification (Table 2), with a recall 0.75/0.81 for MMPose and YOLOv8x-pose, respectively; and a precision of 0.83/0.73. This aiming classification does not yet specify which weapon the person is holding, but is purely driven by the two rules that cover both poses. To test how accurate the difference between these two poses can be estimated, we classified all images with the 'aiming' classification further into 'aiming a handgun' and 'aiming a rifle', for which the results can be seen in Table 2. Here, a difference is observed between the two detectors, since YOLOv8x-pose shows to have a lower performance in distinguishing between the two different poses.

Examples of two images with the keypoint placement and probabilities can be seen in Figure 3. A clear distinction between the two keypoints detection methods is that MMPose always places keypoints in the image, regardless of whether these body parts are visible or not. In the case that these body parts, e.g. the feet, are not in the image, the corresponding keypoints are plotted seemingly randomly and with a very low probability. YOLOv8x-pose only places keypoints that are indeed visible; other keypoints are plotted in the corner at coordinates (0,0) with probability 0. In the next section, we will see more cases of this.

YouTube-GDD – Reduced resolution

YouTube-GDD is a good dataset to test whether our neuro-symbolic approach works, however for our final use case on the DEBELA-dataset, we expect to have some additional challenges. The YouTube-GDD images show the people as the largest, object in the image. In the DEBELA-dataset, the images of people aiming a weapon will be taken from a much larger distance and usually lower resolution. Therefore, we want to test the ability of our method to handle lower resolution data by reducing the size of the YouTube-GDD images. We use only the images with the ground truth classification label 'aiming' and recompute the classifications after scaling the images with factors ranging from 0.1 to 0.01: this is the range in which we see the largest deterioration in performance. This means that with factor 0.1, the image size changes to 128×72 pixels and with factor 0.01 the image size becomes 13×7 pixels. Note that this simple rescaling does not take potential camera blurring and artifacts into account, which will be present in recorded data. Figure 4 shows for both keypoint detection methods the relatively correct and missed 'aiming' classifications.

Table 2. Results of MMPose and YOLOv8x-pose, combined with rule-based classification, for identifying an 'aiming' pose; as well as the results for the classification of 'aiming a handgun' and 'aiming a rifle' poses.

Classification:	**'aiming'**		**'aiming a handgun'**		**'aiming a rifle'**	
	Recall	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**
MMPose	0.75	0.83	0.81	0.74	0.75	0.78
YOLOv8x-pose	0.81	0.73	0.23	0.64	0.49	0.22

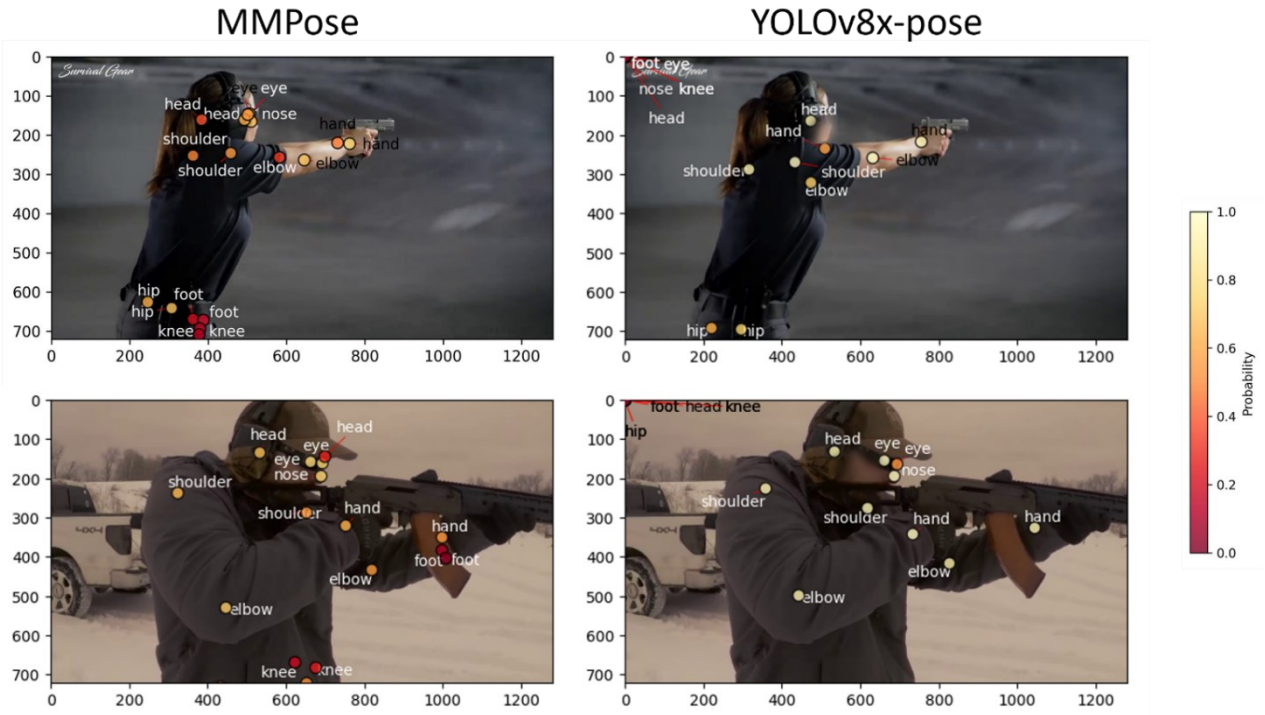

Figure 3. Detected keypoints (coloured points with text labels) with their probability (in colour) for the two keypoint detection models. There is a clear difference between the two methods, especially for the body parts that are not in the image (such as knees and feet). For YOLOv8x-pose, the keypoints plotted in the upper right corner have coordinate (0,0) with probability 0, meaning they are not detected.

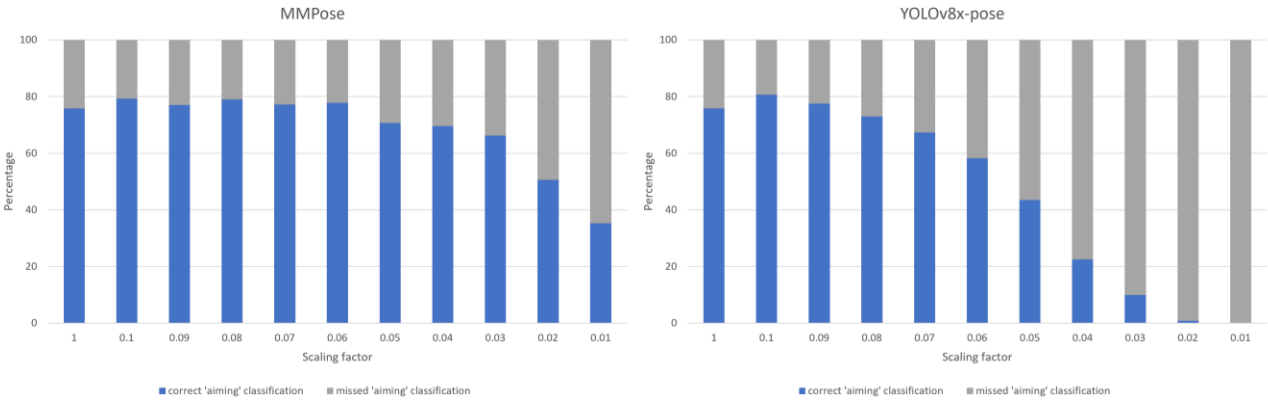

Figure 4. 'Aiming' classification performance for different image resolutions (defined by the scaling of the original image) for both MMPose and YOLOv8x-pose.

Figure 4 shows that MMPose can handle the reduction in resolution better compared to YOLOv8x-pose. Especially after using the reduction factor 0.7, YOLOv8x-pose starts to degrade in performance. When having only 0.01% of the original resolution (scaling factor 0.01), MMPose is able to correctly classify 35%, while for YOLOv8x-pose this is 0%.

To visualize the outcome, a few examples are shown in Figures 5 and 6. The figures show the same images as before, now after scaling both axes with 0.1, the image with the lowest scaling factor that still achieves a correct classification, and the first scaling factor that results in a misclassification; for both keypoint detection methods. Again, we see the difference mentioned before, where MMPose attempts to find a coordinate for unseen keypoints in the image, while YOLOv8x-pose indicates they are not present (and they are plotted at (0, 0)). The general trend that we observe for classification is that, even though MMPose finds keypoints with a very low probability, it still manages to correctly classify the pose; while

YOLOv8x-pose generally has higher keypoint probabilities but 'loses' these keypoints earlier at still higher image resolution. We observe that for correct classification, an estimated minimum number of pixels on target (upper body) is needed. For MMPose this is ~30×30 pixels while for YOLOv8x-pose that is ~50×60 pixels. However, it is noteworthy that this is very dependent on aiming position and point-of-view (front, side, or half-back).

DEBELA-dataset – Qualitative evaluation

The DEBELA-dataset videos are evaluated qualitatively by processing three short videos (videos A, B, and C) where a person is aiming a weapon. We take ten frames per second for each video. Because the weapons are different than in the experiments on the YouTube-GDD, we only classify whether the person is aiming or not. Figure 7 shows the timeline of videos A and B, where each frame is classified as 'aiming' or 'not aiming' by MMPose. A manual estimation of when the aiming position starts and ends is included as well. These qualitative results suggest that MMPose has a good performance, where the 'aiming' classification is starting and ending at approximately the right frames. For Video A, approximately 95% of the 246 frames where the person is aiming are classified correctly. For Video B, this is approximately 99% over 125 frames. On the other hand, YOLOv8x-pose could only correctly classify a handful of frames per video and therefore the timeline classification is omitted from the Figure. Both Video A and B have approximately 250×70 pixels-on-target, which should have allowed for correct classification by both methods. Figures 8 and 9 show the person aiming and not aiming a weapon for both videos.

Figure 10 (top row) shows the timeline results of Video C, which is a harder case for classification since there are approximately 35×35 pixels-on-target for the upper body of the person aiming a weapon. It can be seen that the classification is unreliable and incorrect. To address this, three preprocessing steps are included: (a) cropping the image to guide detection of the person to the correct location in the image, (b) improving image brightness by 50%, and (c) improving brightness and contrast (by using a sharpening kernel to enhance edges and small details). Empirically, best results were obtained when these three steps were combined. Cropping was deemed a necessary step, only altering brightness/contrast did not have the desired result. Increasing brightness after cropping increased the correct classification of 'aiming' from 0% (only cropping) to 14% over 370 frames. Including contrast improvements improved it further to approximately 66% correct classification. Figure 11 shows the original video (2560×2048 pixels) and the corresponding crop (900×600 pixels) that was used for the brightness/contrast preprocessing steps. Figure 12 shows the effect of preprocessing on a frame with an aiming person, and Figure 13 an example of a person not aiming and aiming after preprocessing.

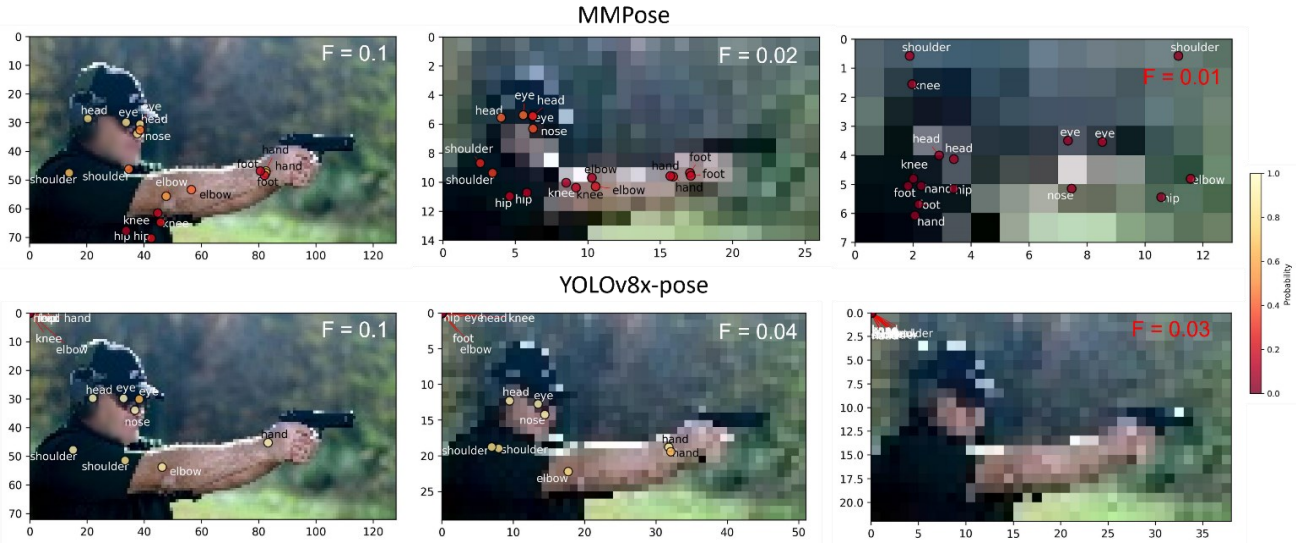

Figure 5. Different image resolutions and the corresponding detected keypoints. F indicates the scaling factor of the original image. For both images, the image with F=0.1 is shown. Additionally, the turning point is demonstrated: the smallest scaling factor with the correct classification is shown (middle) and the largest scaling factor with the incorrect classification (right, scaling factor in red).

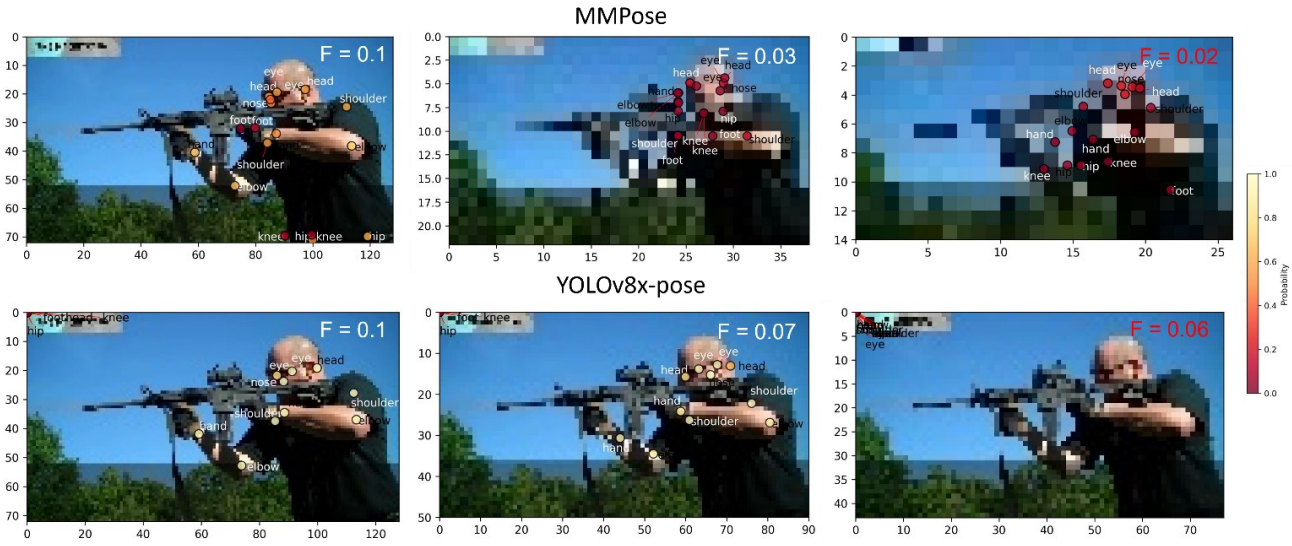

Figure 6. Different image resolutions and the corresponding detected keypoints. F indicates the scaling factor of the original image. For both images, the effect of F=0.1 is shown. Additionally, the turning point is demonstrated: the smallest scaling factor with the correct classification is shown (middle) and the largest scaling factor with the incorrect classification (right, scaling factor in red).

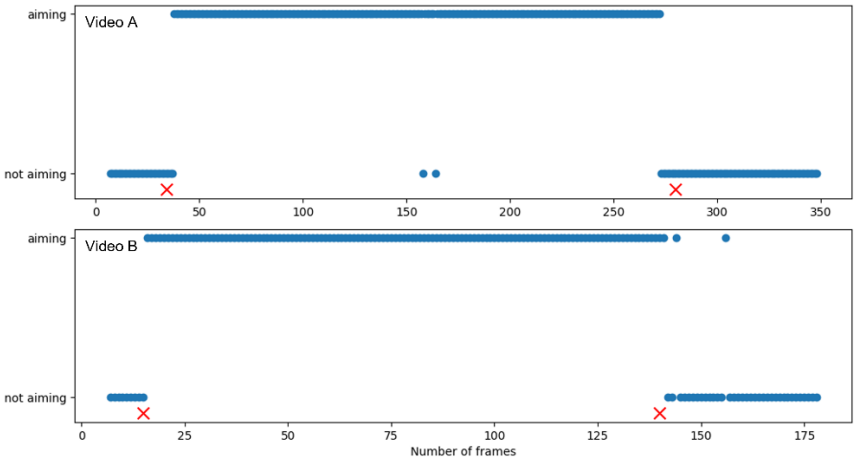

Figure 7. A timeline of two videos from the DEBELA-dataset with the corresponding classification per frame by MMPose. The red crosses indicate the manually annotated start (first cross) and end (second cross) frame in which the person is aiming a weapon.

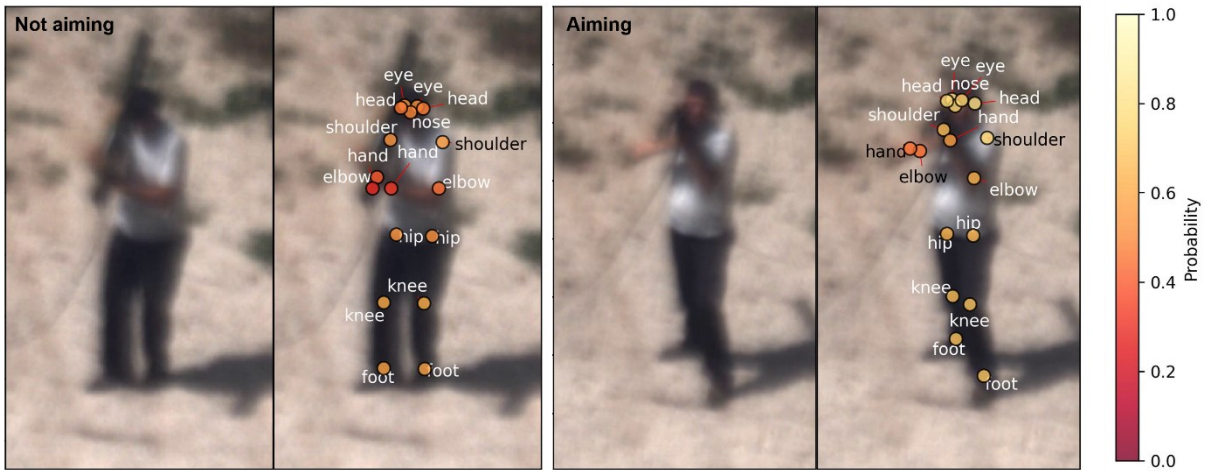

Figure 8. Example frames of Video A, showing a person not aiming (left) and aiming (right) a weapon. The frames are shown with and without keypoints for better visibility. Note that only a 220×370 pixels crop of the full frame is shown.

Figure 9. Example frames of Video B, showing a person not aiming (left) and aiming (right) a weapon. The frames are shown with and without keypoints for better visibility. Note that only a 220×370 pixels crop of the full frame is shown.

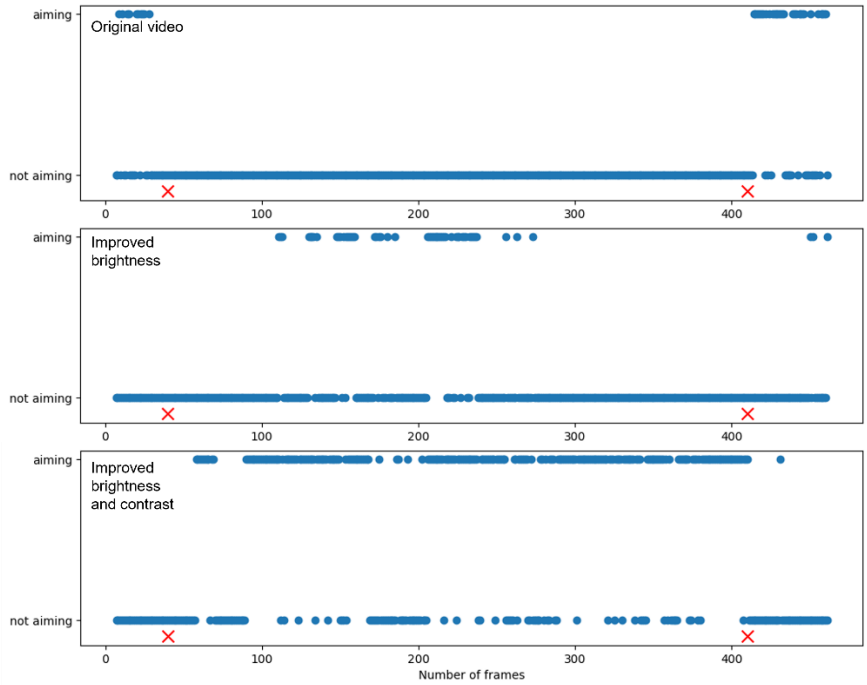

Figure 10. Timeline of Video C and the corresponding classification per frame by MMPose, using three different conditions: the original video (top row), the video with cropping and improved brightness (middle row), and the video with cropping and improved brightness and contrast (bottom row). The latter shows the best performance having the most frames classified correctly.

Figure 11. A frame of Video C with its original image size (left) and the corresponding crop made to direct keypoint detection (right, annotated in yellow on the left image). The original image is 2560×2048 pixels, the crop is 900×600 pixels. The person with a weapon can be seen in the centre, wearing a red shirt, and standing just to the right of a small brick wall.

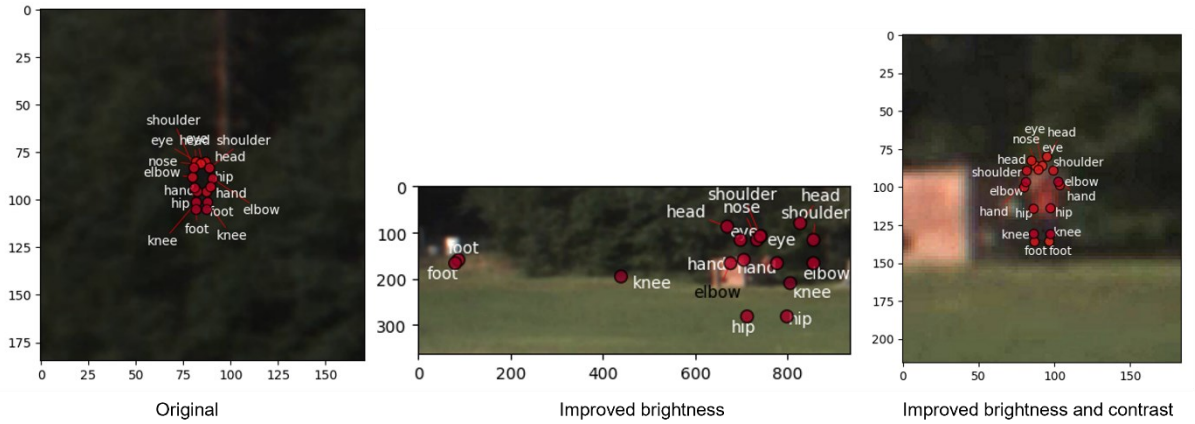

Figure 12. The same video frame as in Figure 11, but then with the detected keypoints overlayed. Left: the frame before preprocessing and the detected keypoints are located randomly in a tree in the background. Middle: after brightness correction, the keypoints are roughly at the correct locations but not precise enough. Right: after brightness and contrast improvement, the keypoints are at the correct location in the image.

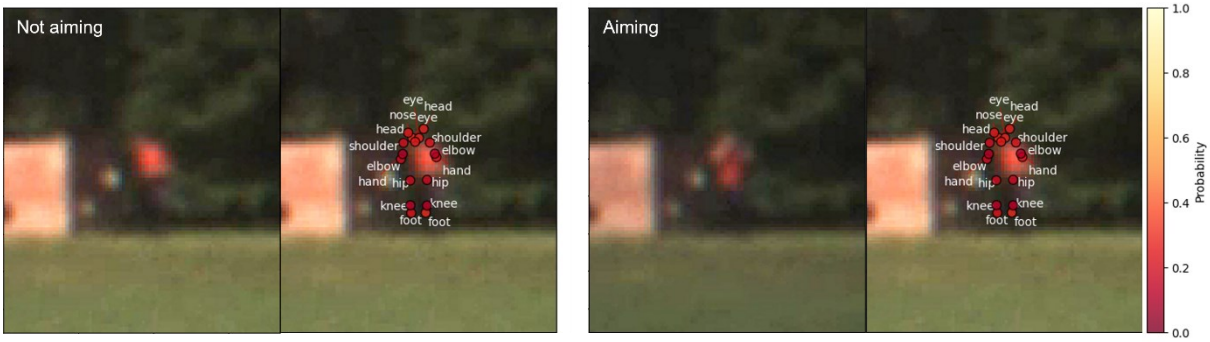

Figure 13. Example frames of Video C after preprocessing the video, showing a person not aiming (left) and aiming (right) a weapon. The frames are shown with and without keypoints for better visibility. Note that only a 180×220 pixels crop of the full frame is shown.

4. DISCUSSION AND CONCLUSION

In this paper, we present a new approach to classify people aiming a weapon, in which AI-based models for body part keypoint detection are combined with the neuro-symbolic framework Scallop. The performance evaluation is both quantitatively as well as qualitatively using two datasets. We used two pre-trained keypoint detection methods (MMPose and YOLOv8x-pose) that can be applied zero-shot, thereby avoiding the need for training. The method is therefore flexible, since new rules can be added easily using the Python interface of Scallop, based on knowledge from domain-experts. We observed that our method shows promising results, however downscaling the image resolution demonstrates the limits of both keypoint detection models and the combination with Scallop. In the case of 13×7 pixels, MMPose is still able to detect 30% of the people aiming a weapon, while YOLOv8x-pose has lost all classification performance.

An advantage of MMPose is that it retains an acceptable performance with reduced image resolution, although the reliability of the keypoints drop indicated by the lower assigned probabilities. A disadvantage is that unseen keypoints are still ‘forced’ into the image, but they can be filtered out with their near-zero probabilities. Owing to the min/max probability provenance of Scallop, the classification framework is still able to correctly classify aiming people even with the lower keypoint probabilities of MMPose. The min/max probability provenance has as key characteristic that it will select the most probable scenario, instead of needing a certain minimum probability threshold. This makes that, even at lower image resolutions, the combination of MMPose and Scallop can classify correctly.

While YOLOv8x-pose has a more correct placement of keypoints, it has considerably degraded performance with lower image spatial resolution. This causes difficulties in our use-case with videos from the DEBELA-dataset, which are acquired with a wider field-of-view. YOLOv8x-pose shows to have a higher keypoint probability, but drops these keypoints quickly after a certain threshold. Whilst Scallop showed to be able to handle low keypoint probabilities and still classify correctly (demonstrated with MMPose), the performance with YOLOv8x-pose reduces quickly since Scallop is not provided with any keypoints as input because YOLOv8x-pose drops these detections. On the other hand, this does make YOLOv8x-pose more reliable, especially when image resolution and the number of pixel-on-target is not too low.

The Python interface for Scallop made it easy and flexible to setup rules. A challenging aspect for defining body pose is the fact that the distance to the people, and therefore their apparent size in the image, can vary. The use of relative distances by using the keypoints for ears and shoulders for normalization partially solved this, with the assumption that the camera is aimed directly at the persons at eyelevel and is not too much higher or lower. For the used datasets, this assumption was correct. Additionally, the optimal scaling factor was found heuristically and one can imagine that this can differ per dataset, per body position, and might be dependent on the keypoints that are chosen. Reducing the image size might also affect these relative distances, since keypoints can be placed differently. As the image resolution becomes smaller, the effect of relative distances can become larger. For every new dataset or image settings, it would be recommended to optimize the distance factors.

Our method showed positive qualitative results for use-case on the DEBELA-dataset, even when having a limited image resolution, owing to some additional image preprocessing. Videos A, B, and C demonstrated the applicability of our method. Even on Video C, which was a hard case to classify, our method shows to be able to classify the pose of the person, even though the results are somewhat unstable over a short range of frames. A possible improvement, especially when dealing with a limited number of pixels-on-target, could be to use a keypoint tracking system over several frames to stabilize the performance. This method could be applied to various security and defence use-cases, by using their own expert knowledge and defining other rules. For the domain of security and defence, methods that can be used with no or limited training data are very useful and valuable. Future work will include other body positions, as well as testing the effects of e.g. occlusion. Both MMPose and YOLOv8x-pose show to have a good estimation for unseen body keypoints based on the rest of the body position, which we observed when e.g. two arms are overlapping in the image. Future work will also aim to improve the classification by combining other imaging modalities, such as infrared, or the use of videos to obtain sequences of poses or activities.

ACKNOWLEDGEMENTS

This work was supported by the European Defence Agency (EDA) CAT B project Detect Before Launch (DEBELA).

REFERENCES

[1] C. Eisele, D. P. Seiffer, E. Sucher, L. Sjöqvist, M. Henriksson, C. Lavigne, R. Domel, P. Déliot, J. Dijk, H. Kuijf and N. Boehrer, "DEBELA – Investigations on potential Detect before Launch technologies," in *SPIE Sensors+Imaging, Electro-optical and Infrared Systems: Technology and Applications XXI*, Edinburgh, 2024.

[2] J. L. S. González, C. Zaccaro, J. A. Álvarez-García, L. M. S. Morillo and F. S. Caparrini, "Real-time gun detection in CCTV: An open problem," *Neural networks,* no. 132, pp. 297-308, 2020.

[3] A. Castillo, S. Tabik, F. Pérez, R. Olmos and F. Herrera, "Brightness guided preprocessing for automatic cold steel weapon detection in surveillance videos with deep learning," *Neurocomputing,* no. 330, pp. 151-161, 2019.

[4] T. Santos, H. Oliveira and A. Cunha, "Systematic review on weapon detection in surveillance footage through deep learning," *Computer Science Review,* vol. 51, no. 100612, 2024.

[5] R. Debnath and M. K. Bhowmik, "Automatic visual gun detection carried by a moving person," *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS),* pp. 208-213, 2020.

[6] J. Ruiz-Santaquiteria, A. Velasco-Mata, N. Vallez, G. Bueno, J. A. Álvarez-García and O. Deniz, "Handgun Detection Using Combined Human Pose and Weapon Appearance," *IEEE Access,* vol. 9, pp. 123815-123826, 2021.

[7] R. Chatterjee and A. Chatterjee, “Pose4gun: a pose-based machine learning approach to detect small firearms from visual media,” *Multimedia Tools and Applications,* vol. 22, no. 83, pp. 62209-62235, 2024.

[8] A. Lamas, S. Tabik, A. C. Montes, F. Pérez-Hernández, J. García, R. Olmos and F. Herrera, “Human pose estimation for mitigating false negatives in weapon detection in video-surveillance,” *Neurocomputing,* vol. 489, pp. 488-503, 2022.

[9] Y. Gu, X. Liao and X. Qin, “YouTube-GDD: A challenging gun detection dataset with rich contextual information,” 2022.

[10] T. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick and P. Dollár, “Microsoft COCO: Common Objects in Context,” 2014. [Online]. Available: https://arxiv.org/abs/1405.0312. [Accessed 21 11 2022].

[11] G. Jocher, A. Chaurasia and J. Qiu, *Ultralytics YOLO (Version 8.0.0),* 2023.

[12] MMPose Contributors, *OpenMMLab Pose Estimation Toolbox and Benchmark,* 2020.

[13] Ultralytics, “Ultralytics YOLO Docs,” 12 November 2023. [Online]. Available: https://docs.ultralytics.com/. [Accessed 22 August 2024].

[14] Ultralytics, “Object Detection,” 12 November 2023. [Online]. Available: https://docs.ultralytics.com/tasks/detect/. [Accessed 22 August 2024].

[15] K. Sun, B. Xiao, D. Liu and J. Wang, “Deep High-Resolution Representation Learning for Human Pose Estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[16] J. Huang, Z. Li, B. Chen, K. Samel, M. Naik, L. Song and X. Si, “Scallop: From probabilistic deductive databases to scalable differentiable reasoning,” in *Advances in Neural Information Processing Systems*, 2021.