Unsupervised domain adaptation in human detection for various human pose

Benoit Lagadec[1], Ekaterina Kostrykina[2]
benoit.lagadec@esifrance.net, ekaterina.kostrykina@etu.univ-cotedazur.fr
[1] ESI, 362 av. du Campon, 06110 Le Cannet, France
[2] Université Côte d'Azur, 28 avenue Valrose 06103 Nice Cedex 2, France

Abstract

In human detection algorithms, some recurrent issues are still challenging. Indeed, in a real application we have a lack of various human poses in our data. For example, in video surveillance it could be a significant issue to forget detecting a human in uncommon pose. We propose a compensation to these less frequent cases. We introduce a detection in rotated view to compensate limited various poses in our dataset. Instead of introducing a generative enhancement for every human body pose, which could be exhausting. Indeed, data annotation has become a difficulty for many applications. Moreover, a simple data augmentation is not efficient to generalize our training. We need a strategy to learn this new poses. So, we will introduce a process to compute pseudo labels in a sequence of rotated views. It avoids to use a clustering algorithm like dbscan[3] or K-means to compute a pseudo label. A summary of theoretical results is described in a first section to illustrate our baseline. First of all, we will illustrate how we process pseudo label on various views by two mechanisms which have been called relaxation and scanning. We introduce a sequence of rotations of original view, to process a detection in a compensated view. In second step, we will learn this distortion created by the various views and construct an approach to distill original knowledge to a more generalized training.

The main improvement of this approach is to propose an alternative solution at dbscan[3] approach which is popular, but which is difficult to use in real applications. We obtain a significant result by improving accuracy in especially various pose datasets.

1. Introduction

Difficulty of human detection could be ordered in two different topics:
The first classes of constraints are based on image quality: Illumination (rain, snow, wind, night, backlight) can perturb the detection. The camera's point of view is still important. Indeed, the different angles of view of the cameras in urban remote surveillance could be a source of issues. We can also notice obstructions, which cannot be neglected in our algorithms. Low image resolution is always present, in fact many security applications must run with not novel cameras.
The second class of constraints focuses more on learning. Indeed, human annotation errors are part of our hypothesis. Moreover, insufficient data or annotations make human detection in image processing more difficult. In this paper, we aim to improve the previous point, specifically addressing the lack of variety in poses. For several years, many neural architectures for object detection have been proposed, and particularly in human detection like: SSD [15], YOLO [17], RetinaNet [19], Fast R-CNN [10], Faster R-CNN [18]. However, these algorithms are not well-adapted for detecting various human poses. In our preliminary study, we identified that a rotated feature map is not the same as the feature map of a rotated image. Based on Fig 3, the clusters are clearly distinct between normal view and various poses. This is why simple data augmentation is insufficient to generalize our training effectively. In the following parts we describe an approach to learn this new poses. Our goal is to learn this perturbation to transform our algorithm and enhance robustness in detecting various views. Our method consists of an adaptation to a particular dataset which is composed of various and rare poses of the person. To make this adaptation, we need to estimate pseudo labels, we introduce an automatic process to estimate or identify localization of lack detection. In Fig 1, the principle of exploration is detailed. Our study is part of a data-driven approach. This work is aimed at improving various poses detection of a given image. Any solution to this problem must tackle a few subproblems. In the following subsections, we describe related work to explain our approach.

1.1. Related work

1.1.1 On detection of various pose

For human detection, a skeleton guide could be used like in [14]. However, these approaches are particularly based on a

CNN most performant to improve step of training with less accurate CNN. We also noticed diverse ways to combine local and global information like it is processed in [20]. To solve this problem of various poses, some approaches take account of perspective estimation as described in [12]: they found a space more adapted by introducing a projection before processing feature map. In [13], person detector combines the distance robust detection scheme with a powerful spatial attention and auto-regressive temporal integration model. Extremely small objects present a significant challenge for existing person detectors [21][1]. In fact, in [1], a zoom operation is proposed instead of our rotation. In this paper, we process our pseudo label in an original manner.

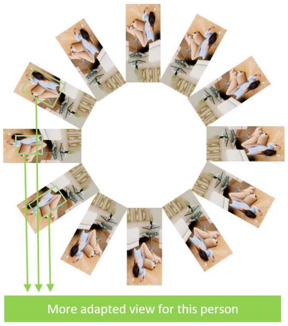

Figure 1. On image with lack of detection, images are rotated to localized candidate bounding box.

1.1.2 On domain adaptation

Many recent papers have been categorized under Unsupervised Domain Adaptation (UDA), which is a component of unsupervised learning. This type of learning is widely applied in various areas, such as re-identification and classification.

1.1.3 On pseudo label

As explained in [3], in a few words, it consists of learning about weak information called pseudo label. Globally, approaches of domain adaptation are based on pre-segmentation of data which is not so accurate [4]. All these constraints make learning/training more complex. To complete unsupervised domain adaptation, a good research direction is detailed in [16] where a real strategy for accuracy of pseudo label is explained. Local and global information are combined to learn pseudo label and to be more precise.

1.1.4 On memory bank

Recently, the contrastive learning combined with the memory bank has been adopted to learn discriminative features for cross-domain.[11]. A hybrid memory to encode all available information from both source and target domains for feature learning is used in [9].

1.1.5 On mutual teacher

Many interests focus on two networks to learn collaboratively with a teaching network and a student network. A mechanism of distillation between theses networks (EMA) is commonly used in [23].

1.1.6 On model

To apply domain adaptation for our use case, we need to justify that a rotated view creates a new domain. Some theoretical aspects have been studied in the following references. The transformation of the input image introduces a perturbation in the learning process. [2].

For model, interesting view of CNN is described in [8]: Each convolutional layer is viewed as a Toeplitz matrix.

Some symmetries are exploited by a generalization of CNNs (Convolutional neural networks) to limit complexity [6]. A steerable CNN is introduced in [5] to transform a CNN in an equivariant by rotation CNN.

In this approach, relied on data improvement: we don't provide a modification of architecture like in [5], we generalize learning to detection in various poses. Our proposed approach consist as a scanning of rotated viewed, where transformation compensates an original pose. By this transform view, detection algorithms are more adapted to human detection.

2. Proposed approach

2.1. Global description

Our main idea is to compensate perturbations in sub-feature maps between a set of various poses of a person and the upright position.

Here, we will process a rotation on image which has a lack of detection. Indeed, we are looking for similarities between local feature map of stand-up person and various poses of person.

In Fig 6, we illustrate our domain adaptation process for these various poses in two steps. We have composed a dataset on Robotflow with a lack of detection in various poses. Our aim is to obtain correct detection, by introducing a sequence of rotations of input image of previous dataset. In consequence, we will introduce a new pseudo label.

By merging the same recommendations obtained in different views, and filtering with a skeleton detection, we have processed a pseudo label. To describe our approach, we introduced an estimation of pseudo label on new poses in a first step of our algorithm.

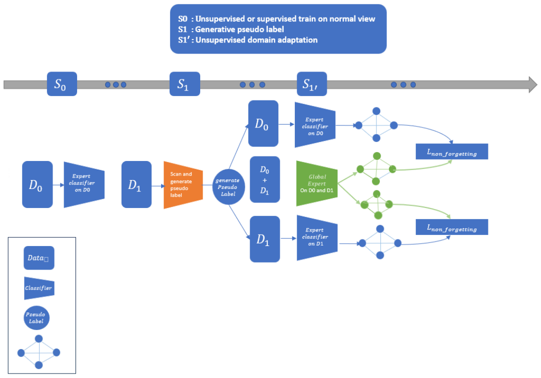

Figure 2. Description of generation of pseudo label and domain adaptation. Adaptation is divided into 2 steps: first to process an estimation of a pseudo label. D1 is our dataset of various poses. D0 is our common dataset which could be learned by supervised or unsupervised manner.

Once we have processed our pseudo label, we use domain adaptation (UDA) to finish our more generalized training while still maintaining our original training on normal views.

In [22], knowledge about two objects is distilled into a global classifier. Here, we distill two views of the same objects. Indeed, rotated views induce a new domain in feature map representation.

3. Theoretical framework

3.1. Overviews

To justify that rotation creates a perturbation in feature map space, we can reference [2]. A simple corollary of the results Cohen and Al is that a rotated view creates a new cluster in the feature map space representation. In our study, particular attention will be focused on integrating various poses into a global classifier.

3.2. Memory loss

To illustrate these results of Cohen and Al, in Fig 3 a new domain appears in this space. We observe that various poses of a person create a distinct feature map compared to the normal pose.

In section 4, we will learn this new representation by introducing a compensated view of our data, which allows to learn and to identify this new feature and this distortion created by a various view. The main difficulty of the following study is to localize the position of lack of detection, to be able to compensate our algorithm for this new view.

In following section, we introduce metrics and loss to make a double distillation of our knowledge. We process alternative data from D0 (normal view) and D1 (various view) datasets: In the first step, we process the feature map from D0 data using a frozen classifier, the feature map is transformed into a predictive distribution P using a softmax function. In the second step, we process the feature map from D0 data with the new classifier to obtain a target distribution Q: Our metric of similarity consists of processing Kullback-Leibler Divergence between these two distributions:

$$\mathcal{L}_{memory} = \mathcal{D}_{KL}(P||Q) \tag{1}$$

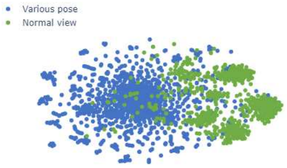

Figure 3. : t-SNE (t-distributed stochastic neighbor embedding) projection: In blue, projection of feature map of stand up person, in green rotated view generated. This representation implies a topology distinction between 2 domains : Domain of normal view and domain of various views.

With two expert classifiers—one from the source domain and the other from the various pose datasets—we manage to distill knowledge to a third classifier, integrating the expertise of these two local classifiers.

So sequentially, we parse in 2 consecutive batchs: data from original dataset and in a second one: data from various pose datasets. Alternatively, we use the frozen classifier to each sub dataset and the previous loss defined to balance our new training.

What we want to learn in this study, is the new domain in blue. We will process this learning as a domain adaptation.

To illustrate perturbations created by rotation views, we have the following repartition of final feature map with a t-distributed stochastic neighbour embedding (t-SNE). Notice than from other representation like Uniform Manifold Approximation and Projection (UMAP), the same distribution appears.

We observed some mixed positions in the feature map projection. Some image projections from the original dataset are clearly included in the new cluster of pseudo labels in blue; we have identified these use cases in our data. It is explained by cropped faces or back views of persons in the original dataset of standing individuals.

4. Approach description

Our detection algorithm is based on a correct pre-train, we use this frozen classifier with a smaller threshold of confidence, we rotate original image and make an inference to see if a compensated view could create a detection. An illustration of distribution is shown in Fig 3. We take as estimation of a pseudo label, a result of detection in a more adapted view. Furthermore, when we have obtained a recommendation of pseudo label, we also filter sub image by a comparison with a key-point detection in a rehearsal view. This last step is considered as post-processing, filtering of

Table 1. Estimation of rotation at different layers of CNN yoloV5.

YOLOV5 layer	0	9	19	29	59
Rotation Estimation	91.92	92.84	96.8	88.15	79.3

pseudo label. We introduce 36 rotations of original image, to see if in candidate view, with frozen classifier, a possible candidate is regular. In the second step of filtering, we merge recommendations of same candidate: If NMS detected a shared area in recommendation of different view: we merge these candidates. For the global dataset, at the end of the generation of pseudo label, we obtained 1177 pseudo labels generated among the 6389 images at the beginning. In the last step of filtering of pseudo label, We make an inference of the key-point detection to see if a part of the skeleton has significant value. This last step improves accuracy of pseudo label, but more details will be given in the experimentation part.

4.1. Domain adaptation to rarely pose dataset

For the domain adaptation to rarely pose, we make a new learning based on this pseudo label. Many approaches of domain adaptation overfed on the target domain like in [23]. Here our approach is designed not to forget the original domain. So in the next section we will introduce a loss to reduce this effect.

By introducing the mechanism of balancing of two losses, we guide the learning to reduce the gap between normal views and various views.

$$\mathcal{L}_{global} = \lambda \mathbb{1}_{target}\mathcal{L}_{memorytarget} + \mathbb{1}_{source}\mathcal{L}_{memorysource} \qquad (2)$$

$\mathbb{1}$ is a indicator function. λ is a hyperparameter where tunning is detailed in Fig 4. Then we learn new various poses with pseudo label generation, without forgetting to learn our supervised dataset. By this way, we have processed by an unsupervised manner, without no more label in our new dataset.

5. Experimentation

5.1. Validity of theoretical results

In this section we have validated few theoretical results described in [6][7] : a rotating view of image processed by a CNN, produce a rotated feature map in only case of $\pi/2$. To check this, we estimate the rotation by maximum cross correlation (MCL). In table 1 results of estimation: We measure at different output of CNN, the rotation. We artificially rotated the original image by 90 degrees with the center of the image as the center of rotation. So, the ground truth rotation is 90 degrees. In the table 1 we evaluate on a set of 10 images the estimation of rotation.

Table 2. Confident threshold for generation of lacks dataset.

Threshold	0.1	0.2	0.3	0.4	0.5
Images in lack dataset	89	198	362	575	959

We also conducted experimentations for other angle by evaluating matrix of cross correlation. It confirm theoretical results of [6]. We are not able to observe a real conservation of structure of feature map for other angle. So we will be urged to scan different angle to learn the different feature map corresponding to a various view of someone.

5.2. Dataset/Settings

5.2.1 Normal views

For our application, we have composed a dataset of 19331 images: D_0. Resolution of our images is in range: QCIF and 9CIF . We have constructed our dataset with different characteristics. We identify 5 views. First one is called human views with 9952 images, where humans are in foreground. Another one is called aerian view, with majority of drone views, this subset has 5353 images. The next view consists of individuals primarily positioned in the depth of the image. It is composed of 1,101 images. We have a subset of 1,477 thermal images and a portion of 1,448 images taken at night.

5.2.2 Various views

We composed a dataset of various poses : D_1, sourced from Roboflow, consisting of 2,647 images from sports events. Data can be accessed here. This dataset has the particularity to have at least one person present in image.

5.3. Pseudo label generation

We process frozen classifier (trained on D_0) to measure the following lack of detection with different confident threshold on all images of D_1 . In the backbone, we use CNN: yoloV5-X. For this task of creation of lack dataset, more confident threshold is low, less permissive we are for lack detection.

We obtain the following parameter for different threshold. First parameter described is set to $iou_{threshold} = 0.40$. Second one to merge bounding box is set to $iou_{thresholdNMS} = 0.26$. Last one, for reducing noise effect of bounding box is $iou_{thresholdNMSall} = 0.40$. Model confidence for relaxation process is set to $model_{conf} = 0.15$.

Regarding this repartition, we made a compromise and chose a threshold of 0.5 to have enough pseudo labels for our domain adaptation. So, our new dataset of various poses (D1) was composed of 959 images. Enhance we process our algorithm of generating: For each image, we introduce sequences of rotations of 10 degrees. We make the same

Table 3. We describe Number of image process during our step of generation/filtering.

Dataset	Noisy	No noisy
Number image original dataset	6389	1292
Number Image with Lack detection	727	232
Image generated with keypoint filtering	412	184

Table 4. Measure of forgetting effects. In the following table, evolution of precision and map criteria. For mixed view, statistically we have one image of D1 per batch.

	D0: Source			**D1: Target**		
Test on Train on	Precision	Recall	Map	Precision	Recall	Map
Expert on Source	0.907	0.882	**0.938**	0.592	0.427	0.413
Expert on Target	0.382	0.602	0.386	0.638	0.631	**0.614**
Source and target mixed	0.938	0.855	**0.933**	0.561	0.513	0.526
Ours with $\lambda = 0.5$	0.935	0.827	**0.919**	0.694	0.583	**0.645**

number of inferences per rotated image, and merge recommendations in double.
In table 3 we add a skeleton filter, we solved 596 images out of 959, there might be 1-3 pseudo-labels for each image.
In Fig 3, we use t-SNE projection to display the repartition between the feature map of standing individuals and the pseudo label feature map. If we refer to Table 3, in the first column, we observe that 57 percent of lack dataset which is partially solved before our domain adaptation. In the second column, with a dataset without noise, we are near 79 percent.

Our main CNN is based on YOLOv5, which has shown good accuracy on various human detection datasets. Future work will involve transposing our approach to the YOLOvX architecture. In the first step, we will train our network in a supervised manner using a common dataset.

This selected image will be integrated in the next section to improve various poses in a domain adaptation.

5.4. Unsupervised domain adaptation

Many domain adaptation approaches overfit to the target domain. In this section, our aim is to generalize our detection algorithms while minimizing forgetting effects. We refer to our dataset of normal views as the source domain and our new dataset of various poses as the target domain.

5.4.1 Results with memory metric

In the first batch with the frozen classifier, the classifier is obtained from our supervised training. We also process the same D0 data with our new classifier, resulting in a new feature map. We introduce a comparison using Kullback-Leibler loss to maintain a level of consistency between these two weights. In the second batch, we take new data from D1 and apply the same similarity measures with another expert frozen classifier. To resume our approach, the following algorithm is detailed in Alg 1.

Algorithm 1 Distillation :

```
procedure GENERATE A THIRD CLASSIFIER( D0, D1)
     On a batch on source
    featuremap_source ← inferenceExpert(source_image)
    featuremap_adapted ← inferenceAdapted(source_image)
    $\rightarrow \mathbb{1}_{source}\mathcal{L}_{memorysource}$
     On a batch on target
    featuremap_target ← inferenceExpert(target_image)
    featuremap_adapted ← inferenceAdapted(target_image)
    $\rightarrow \mathbb{1}_{target}\mathcal{L}_{memorytarget}$
end procedure
```

5.4.2 Discussion

In Table 4, we observe a significant forgetting effect in the second line of training. Additionally, integrating our new pseudo labels as samples from our original dataset is not optimal for the target domain (third line). The second line indicates that specialized training is more effective for the rotated data. Therefore, an adaptation step could help find a better compromise between these two domains. Our limit bound for the normal view using the MAP criteria is 0.938.

- Between the second and third lines, we achieve better integration of various views, but it is still not optimal: our limit bound for the various views using MAP is 0.614.

- Between the third and fourth lines, we see improved precision, recall, and MAP on the target domain. However, for the MAP criteria on the source domain, this is not the case, as we have a less accurate pseudo label due to the generation step, which explains the lower recall.

- Between the first and fourth lines, we limit the forgetting effect on the source domain while enhancing the classifier's expertise on the target domain. The fact that the MAP criteria outperform our local expert could be attributed to the overlap between clusters in the source and target domains, even though they are distinct.

On validation metrix on target, we split our new dataset randomly. To be compared with supervised ways, we decided to not mix our source and our target dataset. In real case, we cannot evaluate the frequency of apparition of this various view.

To summarize, we observe a substantial improvement in various views, with a 23 percent increase compared to the initial view. Moreover, by introducing a balanced loss, we prevent forgetting the initial source domain, achieving nearly the same score on the source with only a 2.

5.4.3 Hyperparameter tunning

In final study we resume the results to find the best value for hyperparameter λ on validation set:

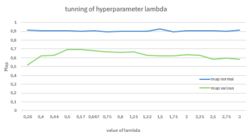

Figure 4. We test different value of hyperparameter: λ which balanced our two local loss.

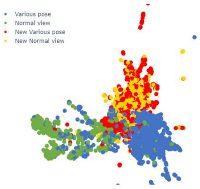

Figure 5. Results of repartition after distillation. In yellow and red the new repartition of normal and various pose. Distance between clusters has decreased.

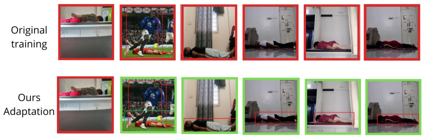

Figure 6. On the first line, an example of lack dataset. On the second line, the results after distillation.

6. Conclusion

We have constructed an unsupervised algorithm with a significant improvement in accuracy. We adapted our approach to the case of rotated views. Our method is designed to be used in real applications, particularly in video surveillance. Demonstrating significant improvement with limited new data, a further step is to add more various poses, though data collection is challenging. An important way to improve this approach is to propose compensation for rotation during inference in our proposed method. By balancing our data, we reduce the forgetting effect introduced by learning sequences of datasets. A more theoretical approach would be to modify the layers of the CNN to maintain invariance by rotation. This could be a future work to compare these two approaches.

References

[1] F.C. Akyon, S.O. Altinuc, and A. Temizel. Slicing aided hyper inférence and fine-tuning for small object detection. *ICIP*, 2022. 2

[2] A. Azulay and Y. Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning*, 2019. 2, 3

[3] H. Chen, B. Lagadec, and F. Bremond. Ice (instance contrastive encoding): Inter-instance contrastive encoding for unsupervised person re-identification. *ICCV*, 2021. 1, 2

[4] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, and F. Bremond. Learning invariance from generated variance for unsupervised person re-identification. *TPAMI*, 2022. 2

[5] S. Cohen and M. Welling. Osteerable cnns. *ICLR*, 2017. 2

[6] T.S. Cohen and M. Welling. Group equivariant convolutional networks. *ICML*, 2016. 2, 4

[7] T.S. Cohen, M. Geiger, J. Köhler, and M. Welling. Spherical cnns. *ICLR*, 2018. 4

[8] B. Delattre, Q. Barthélemy, A. Araujo, and A. Allauzen. Efficient bound of lipschitz constant for convolutional layers by gram iteration. *ICML*, 2023. 2

[9] Y. Ge, F. Zhu, D. Chen, R. Zhao, and H. Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *NIPS*, 2020. 2

[10] R. Girshick. Fast r-cnn. *ICCV*, 2015. 1

[11] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. *CVPR*, 2020. 2

[12] Y. Hou, L. Zheng, and S. Gould. Multiview detection with feature perspective transformation. *ECCV*, 2020. 2

[13] D. Jia, A. Hermans, and B. Leibe. A spatial-attention and auto-regressive model for person detection in 2d range data. *CVPR*, 2004. 2

[14] X. Ju, A. Zeng, C. Zhao, J. Wang, L. Zhang, and Q Xu. Humansd: A native skeleton-guided diffusion model for human image generation. *ICCV*, 2023. 1

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, and A.C. Berg. Ssd: Single shot multibox detector. *ECCV*, 2016. 1

[16] L.Long, T.Xiao, C.Haoang, and Z. Xiang. Multiscale knowledge distillation for unsupervised person reidentification. *ICCV*, 2022. 2

[17] J. Redmon, S.K. Divvala, and R.B. Girshick. You only look once: Uni-fied, real-time object detection. *CVPR*, 2016. 1

[18] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NIPS*, 2016. 1

[19] Lin T.Y. and R. Goyal P.and Girshick. Focal loss for dense object detection. *ICCV*, 2017. 1

[20] J. Yang, A. Zeng, S. Liu, F. Li, R. Zhang, and L. Zhang. Explicit box detection unifies end-to-end multi-person pose estimation. *ICLR*, 2023. 2

[21] X. Yu Yuqi and Gong Nan Jiang Qixiang Ye Zhenjun Han. Scale match for tiny person detection. *WACV*, 2020. 2

[22] H. Zhang, F. Mao, M. Xue, G. Fang, Z. Feng, J. Song, and M. Song. Knowledge amalgamation for object detection with transformers. *TIP*, 2023. 3

[23] X. Zhang, Shijian, H. Gong, Z. Luo, and Ming Liu. Adversarial-based mutual learning network for online knowledge distillation. In *ECCV*, 2020. 2, 4