Collaborative training of far infrared and visible models for human detection

Paul Blondel, Alex Potelle, Claude Pégard, Rogelio Lozano

Int. J. Simul. Multidisci. Des. Optim. **10**, A15 (2019)

https://doi.org/10.1051/smdo/2019016

IJSMDO
Available online at:
https://www.ijsmdo.org

RESEARCH ARTICLE OPEN ACCESS

Collaborative training of far infrared and visible models for human detection

Paul Blondel[1,2,*], Alex Potelle[1,2], Claude Pégard[1,2], and Rogelio Lozano[1,2]

[1] Université Picardie Jules-Verne, 80000 Amiens, France
[2] Université Technologique de Compiègne, 60200 Compiègne, France

Received: 15 May 2019 / Accepted: 27 August 2019

Abstract. This paper is about the collaborative training of a far infrared and a visible spectrum human detector; the idea is to use the strengths of one detector to fill the weaknesses of the other detector and vice versa. At first infrared and visible human detectors are pre-trained using initial training datasets. Then, the detectors are used to collect as many detections as possible. The validity of each detection is tested using a low-level criteria based on an objectness measure. New training data are generated in a coupled way based on these detections and thus reinforce both the infrared and the visible human detectors in the same time. In this paper, we showed that this semi-supervised approach can significantly improve the performance of the detectors. This approach is a good solution to generate infrared training data, this kind of data being rarely available in the community.

1 Introduction

Human detection is still a challenging problem despite the recent advances in the field. There is still room for improvement to make human detection more robust. The classic approaches to enhance the detection often consist of designing better features and/or choosing better classification methods. However, in many cases the performance of a detector can be simply improved by adding new data to the training dataset. New training data can be generated by crossing the detection results of different human detectors, such as: a visible and a far-infrared spectrum human detectors.

1.1 Human detection in the visible spectrum

The best performing human detectors use trained models. At first, a model (or classifier) is trained with training images, at run-time, the trained model is inferred to decide whether or not there are detections. The Histogram of Oriented Gradients (HOG) detector of Dalal and Triggs is one of the first approach to reach significant human detection performance using a trained model [1]. Since this moment, researchers worldwide continued to improve the speed and performance of human detectors. Algorithms based on the Integral Channel Features (ICF) detector of Dollàr et al. [2] over-performed afterwards the HOG detector in terms of speed and detection performance. With these detectors, integral images are used as features and a Boosting soft-cascade model is used for classification. Soft-cascade classifiers are fast, due to the use of a rejection trace. Amongst these ICF-based detectors: the Fastest Detector in the West (FPDW) [3] approximates the features in the image pyramid so it is not needed to compute a dense image pyramid (thus it is less time-consuming), whilst the more recent and faster Aggregate Channel Features (ACF) detector performs pixel look-ups in aggregated image channels [3]. Researchers recently considerably improved the accuracy of object detection using deep-learning techniques: the faster-RCNN and the Yolo detectors are very famous detector based on deep-learning [4,5]. Because the proposed approach of this paper is agnostic to the type of detector and because deep-learning approaches require a lot of training data to work well we chose, instead, to demonstrate our approach using the ACF detector as a baseline.

1.2 Human detection in the infrared spectrum

Most of the infrared spectrum human detectors have a design very similar to visible spectrum detectors. Indeed, researchers previously compared different configurations of human detectors for the task of detecting people in infrared images and they suggested that there is no need to invent and use radically different methods to perform well in this modality [6]. Different kinds of discrete features have been proposed by Olmeda et al. to get a robust infrared spectrum descriptor: the histogram of phase congruency orientation [7]. The sensitivity to changes of contrast is less important

* e-mail: p.blondel@net.estia.fr

when using the phase congruency. Thus this performs well on infrared images subject to abrupt changes of contrast [8]. But, it is extremely computationally intensive to compute the phase congruency of an image. It is more convenient and faster to use simpler and still very efficient descriptors such as the features of the HOG detector or the ICF-based detectors. In their work, Brehar et al. proposed to combine a search space reduction technique with a modified ACF detector to reach near real-time detection [9].

1.3 Collaborative training

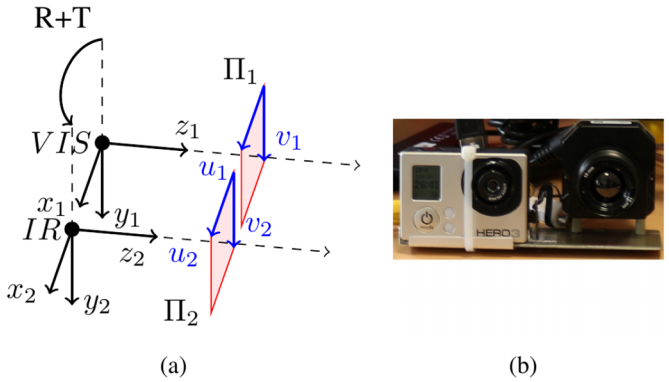
The collaboration of the trainings of the infrared and the visible spectrum human detectors can be done using a co-training approach. Co-training was first proposed by Blum and Mitchell as a semi-supervised approach to improve the training of different classifiers [10]. The classifiers must treat different but complementary "views" of the scene. Levin et al. proposed a co-training approach using two differently designed detectors associated to one visible camera [11]. Having various "views" of the scene has a bigger impact on the co-training. This can be achieved by using different modalities (as is proposed in this paper) or by having different physical points of view of the scene, as Roth et al. proposed [12]. Roth et al. used several cameras in order to obtain distinctly very different points of view of the scene. To the best of our knowledge, co-training has not yet been used with a far infrared human detector before.

1.4 Content of the paper

In this work a far infrared and a visible human detector work together at training; the training of the detectors is improved in a co-training procedure using low-level thermal information and detection results to extract new training samples. The infrared and the visible information are synchronized in order to extract pairs of multimodal training samples. In this paper, the two detectors that we are using are: the ACF detector of Dollàr et al. [3] and the adapted ACF detector of Brehar et al. [9]. For the purpose of this research work, we built our own heterogeneous stereoscopic system which is composed of a far-infrared camera and a visible camera. In the Section 2 the design of our stereoscopic system is described as well as the methods that we used in order to synchronize the frames coming from the two modalities together, in the Section 3 the ACF detector of Dollàr et al. is described in details. Section 4 describes our collaborative approach of training the two models together and the Section 5 is about the tests and results of our proposed approach.

2 Hardware

At first, it is required to align the visible and the far-infrared images coming from our two cameras. We have built for this purpose a stereoscopic system made of a Flir tau 2 camera (far infrared spectrum camera) and a GoPro camera (visible spectrum camera). These cameras are placed one next to the other as illustrated in Figure 1.



(a) (b)

Fig. 1. (a) The layout of our stereo acquisition system. VIS and IR are respectively the visible and the infrared camera, $\prod_1$ and $\prod_2$ are the image planes of respectively camera VIS and IR. R: rotation matrix from camera VIS to camera IR, T: translation matrix from camera VIS to camera IR. (b) Picture of our stereoscopic system.

Aligning spatially the images of the two cameras is not enough. It is important to also align temporally the frames. Indeed human patterns should be present at the same places in our images – and – appear at the sames moments in our video streams.

Two important steps are thus required:

(1) *The temporal synchronization of the frames*: there might be a temporal shift between the visible and the far-infrared video streams when grabbing the images from them two cameras. The cause of this shift can be due to the fact that the two cameras have different acquisition frequencies. Besides, cameras often suffer from a slight delay before starting to film. Different cameras might have different delays.

In order to solve this issue one can usually consider two approaches: (i) triggering the acquisition of the other camera using the trigger output of the other camera, or (ii) fix the shift afterwards with a reference top. But for this work, we managed to obtain a near perfect temporal synchronization using an identical acquisition pipeline for the two cameras. The two cameras are converted from analogical to digital signals using grabber devices having the exact same specifications. We reached a maximum temporal shift of one frame with this approach. This shift was small enough not to have negative repercussions on our experiments.

(2) *The spatial synchronization of the frames*: Krotosky et al. proposed four different approaches for the spatial synchronization of the frames of a parallel-axis stereoscopic system such as the one used in our experiments [13]. These four synchronization approaches are: (i) the global image registration, (ii) the stereo geometric registration, (iii) the partial image registration (using ROIs) and (iv) the infinite homography registration. (i) The synchronization of the objects contained in the filmed scene can be done for a specific depth of the scene, this is the global image registration of the frames. Because this is a global approach the synchronization might not be as accurate for the whole image. (ii) With this approach we use a third and new information: the depth cue. This cue can be used to

Fig. 2. Examples of synchronized pairs of visible and far-infrared images using the infinite homography registration. People close to the stereoscopic system are not well synchronized, people far from it are well synchronized.

geometrically register objects depending on where they are in the depth of the scene, this is the stereo geometric registration. Stereo registration is usually performed using two identical cameras configured as a parallel-axis stereoscopic system. (iii) With partial image ROI registration, only some areas (or ROIs) in the filmed scene are registered and geometric transformations are used so that the objects in the ROIs are synchronized. (iv) The last approach use homography transformations so that objects located at very long distance ("infinity") are matched each other with the assumption that, at infinity, the objects of the filmed scene perfectly overlap in the two modalities.

We decided to choose the infinite homography approach (Fig. 2). There are three reasons why this approach is well suited for our experiments: (i) for human detection applications this is not possible to make any assumption about where will be the persons to detect in the scene, (ii) we thought it was not a wise idea to add an extra visible camera in our setting in order to benefit from the depth cue and (iii) we wanted to avoid a computationally intensive synchronization approach to get a fast co-training of the two models.

As mentioned above we make the assumption that the persons we want to detect are located at a long distance from the parallel-axis stereoscopic system with the infinite homography registration approach. In order to make valid this assumption we simply have to make sure that the baseline between the two cameras is very small when compared to the distance from the cameras to the persons we want to detect in the filmed scene [13]. In our case we have to make sure the cameras are close enough to each other and the persons to detect are several meters away. We need several parameters to compute the infinite homography transformation. We need: a rotation matrix R describing the rotation transformation from the VIS referential to the IR camera referential, and the matrices of intrinsic parameters K_1 for the camera VIS and K_2 for the camera IR.

With these three parameters we can compute the infinite homography matrix as follows:

$$H_\infty = K_2 \times R \times K_1^{-1}. \tag{1}$$

The pixels of the VIS camera can be projected onto the image plane of IR camera with this mathematical formula:

$$\begin{pmatrix} u_2 \\ v_2 \\ w_2 \end{pmatrix} = H_\infty \times \begin{pmatrix} u_1 \\ v_1 \\ 1 \end{pmatrix}. \tag{2}$$

The matrices of intrinsic parameters K_1 and K_2 as well as the rotation matrix R have been obtained using the Bouguet's Matlab toolbox [14]. This toolbox requires a large number of image pairs filming the same referential. Using the toolbox the same corner points must be selected for both the images. We chose as a referential a checkerboard made of black and white squares. We heat up this checkerboard in order to better perceive and select the same corners points but in infrared images. We collected numerous pairs of visible and infrared images to perform the estimation of the intrinsic parameters and the rotation matrix.

3 The basis detector: the aggregate channel features detector

The ACF detector of Dollàr et al. [3] extracts features on ten different image channels from a previously filtered

image. These channels are: the three components of the LUV color space, a normalized magnitude channel and six other channels corresponding to six orientation bins of the HOG descriptor used with the HOG detector. For each channel the image is divided into 4×4 adjacent pixel blocks. The pixel blocks are then aggregated giving an aggregated version of each image channel. The aggregated channels are filtered again [3]. Visual features are obtained just by pixel lookups inside these 10 now filtered aggregated channels. This transformation is performed at training and at during the detection.

An infrared version of the ACF detector has been proposed by Brehar et al. [9]. With this version, only 8 channels are used. The three channels of the LUV color space are just replaced by one grayscale channel. This detector will be named the IR-ACF detector is the rest of the article. In this article we do not use the proposed search space reduction used conjointly with the IR-ACF detector [9].

Both the ACF detector of Dollàr et al. and the IR-ACF detector of Brehar et al. use a Boosting soft-cascade model. In their work, this model is a serial combination of 2048 depth-2 decision tree weak-classifiers. During the training, and for each of these week-classifiers, we just look for the pixels in the filtered aggregated channels that best separate the training positive samples (images of persons) from the negative training samples (image of everything except persons) [3]. After having trained one weak-classifier, the training samples get different weights in order to influence the training and thus optimize the separation between the positive and negative classes. During the detection all the weak-classifiers are inferred one after the other (as long as the accumulated score is still above the rejection trace [2]). If all weak-classifiers are passed we consider that there has been a detection, otherwise we consider there has been no detection.

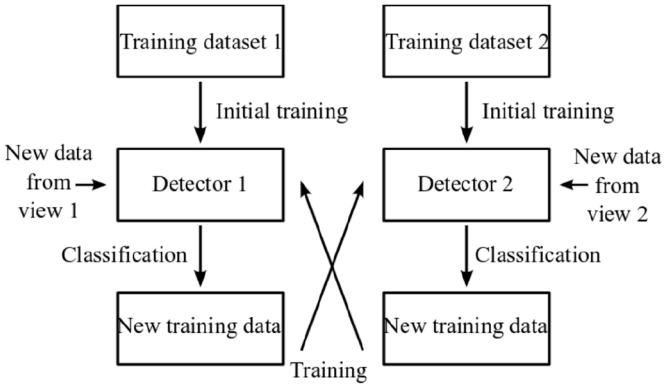


Fig. 3. Co-training's principle.

4 Collaborative training

A mutual and semi-supervised improvement of the training of the detectors is made possible thanks to co-training. The idea behind co-training is quite straightforward: two (or more) classifiers reinforce each other by learning from the results of each other (Fig. 3). Two assumptions are required: the conditional independence of the classifiers and the existence of initial weak classification rules [10]. Note that it helps us acquiring new training data that could be difficult to find otherwise (to form a competitive infrared training dataset).

Our co-training procedure is based on two steps: (1) candidate pairs of regions of interest (ROIs) are selected based on their thermal objectness scores, and (2) the updated training datasets are filtered using a boosted noise filter to eliminate eventual mislabeled samples [15] (see Fig. 4). The procedure can be repeated n times. Note that new negative samples are randomly resampled after each addition of positive samples in order to ensure a good balance between the positive and the negative samples in the datasets.

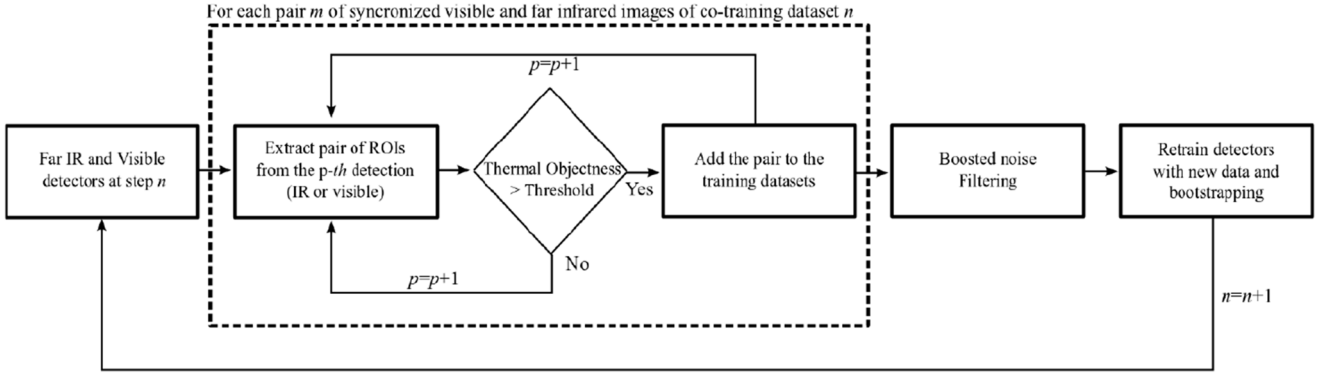


Fig. 4. Diagram of our co-training procedure. For the m-th image pair, an infrared/visible pair of regions of interest (ROIs) is extracted for each detection p. A score of thermal objectness is computed on the infrared ROI. If it is bigger than a threshold, the pair is injected in the training datasets as two new positive samples. If it is lower than a threshold, the pair is rejected. The new training datasets are filtered to eliminate eventual mislabeled samples. Finally, the detectors are retrained with the new data. It can be repeated n times for iterative improvements.

Fig. 5. Examples of detections obtained on visible images and far-infrared images. We can see that the strengths and weaknesses of the far-infrared and the visible detectors are different.

4.1 Measure of the thermal objectness

The first step consists of selecting candidate pairs of training samples in the detections using a measure of thermal objectness (examples of detections are given in Fig. 5). The score proposed in this paper for measuring the thermal objectness is based on the Edge Box (EB) score of Zitnickal [16]. It has been observed that performing an edge-based analysis in the far infrared spectrum is more convenient than in the visible spectrum. In the visible spectrum, there are more ownership contour ambiguities between the different objects of the scene and/or the surrounding. Indeed, contours in the far infrared spectrum correspond to thermal separations between warm and cooler objects of the scene. (1) The original EB score: the EB score based on the observation that the number of fully contained contours is indicative of the likelihood of the box to contain an object [16]. It is computed as follows: at first the edge map is computed and contours are formed by grouping edge pixels of similar orientation values with small edge pixel groups (or contours) being merged to make bigger ones. An affinity measure is computed for each pair of contours: the affinity between two contours sk and sk0 is high if the angle between the groups' means ($\theta_{kk'}$) is similar to the groups' orientations (θ_k and $\theta_{k'}$) (Eq. (3)). γ is the affinity sensitivity, it is generally set to 2. The affinity is equal to zero if the contours are separated by more than 2 pixels.

$$a(s_k, s_{k'}) = |\cos(\theta_k - \theta_{kk'})\cos(\theta_{k'} - \theta_{kk'})|^{\gamma}. \qquad (3)$$

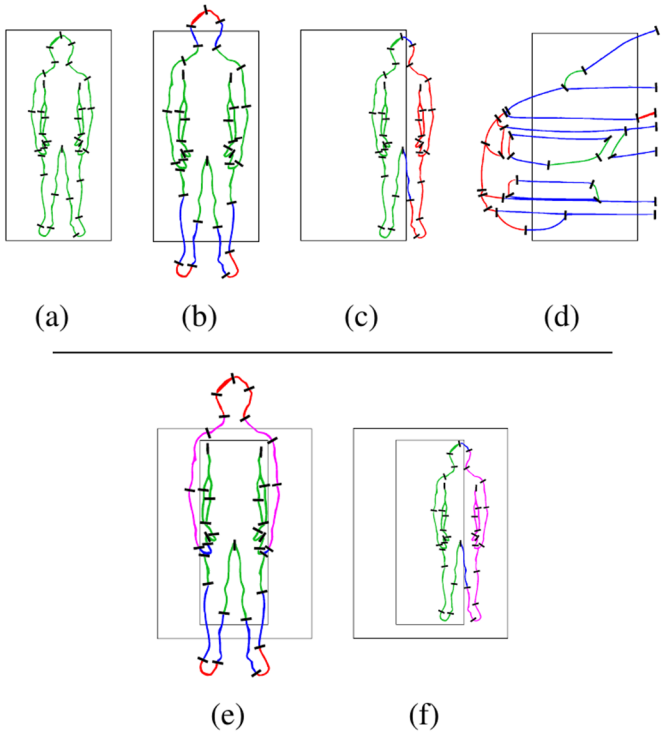


Fig. 6. Above: contours used for the EB score. In green: internal contours, in blue: intersecting contours and in red: external contours. From (a) to (d): the EB score decreases (the number of internal contours decreases and the number of traversing contours increases). Below: the additional contours used for the CAO score. In pink: affiliated external contours. In (e) and (f): as the number of affiliated external contours increases the CAO score decreases.

There are three sets of contours: the set of external contours, the set of traversing contours (S_t) and the set of internal contours (S_i) (see the above part of Fig. 6). External and traversing contours have a null contribution to the final score. For each contour s_k a weight $w_k \in [0, 1]$ is computed: $w_k = 1$ for a pure internal contour and wk decreases if s_k is connected to a traversing contour (Eq. (4)). T is an ordered path group beginning for some

$t_1 \in S_t$ and ending for $t_T = s_k \in S_i$.

$$w(s_k) = 1.0 - \Pi_{j=1}^{|T|-1} a(t_j, t_{j+1}). \quad (4)$$

The final score H is defined by equation (5) m_k is the sum of the magnitudes of the contour s_k, and b_w and b_h are, respectively, the width and the height of the original box. (2) Our Centered and Anti-overflow Objectness (CAO) in infrared images: the EB score may be relatively high for a non-centered object or an object overflowing the box. This could lead to a detection being considered as an exploitable positive training sample due to its high score, whereas the low-level information does not match a true positive detection (human being detection). Centered and Anti-overflow Objectness (CAO) avoids this drawback. Indeed, CAO scores computed on infrared images are high for centered, warm objects of the scene.

$$H = \frac{\sum_{k=0}^{\mathrm{card}(S_i)} w(s_k) m_k}{2(b_w + b_h)^{\kappa}}. \quad (5)$$

$$H' = H - \frac{\sum_{k=0}^{\mathrm{card}(S_e)} w(s_k) m_k}{2(b_w + b_h)^{\kappa}}. \quad (6)$$

The CAO score is computed as follows: two concentric areas of analysis are considered (see the below part of Fig. 6). S_t is now the set of the traversing contours of the inner box and S_i is the set of the internal contours of the inner box. S_e is the set of the affiliated external contours: contours contained between the outer box's border and the inner box's border and having a non-null affinity path to one of the traversing contours. The final score H_0 (Eq. (6)) is simply the EB score obtained for the internal contours minus the EB score of the affiliated external contours.

4.2 Boosted noise filtering

A second step is often required to filter mislabeled training samples. The low-level thermal objectness step can indeed sometimes fail to reject false positives. The boosted noise filter has the ability to identify mislabeled training data [15]. Boosting is very sensitive to mislabeled data and this property is used to identify the mislabeled data [15]. Each training sample has an associated noise counter with training samples having a high noise count likely being mislabeled cases.

5 Results

5.1 Mislabeled samples

This part is about testing the filtering ability of the boosted noise filter and thus its relevance in the pipeline. The filtering ability has been tested on the INRIA dataset [1] for different percentages of mislabeled samples (Fig. 7); we arbitrary chose the visible detector for this test. The percentage was increased with a step of 2.5%. Each time we filtered the samples with the worst scores [15] and we removed a number of samples equal to the number of mislabeled samples present in the dataset. For each step, we took the average filtering results obtained for 10 different randomly produced batches of mislabeled samples. We can see in Figure 7 that the percentage of retrieved mislabeled samples is about 90%, from 0% to 20% of mislabeled samples. The filtering becomes very inefficient from 20% and upwards of mislabeled samples. Figure 8 shows some filtered mislabeled cases we have got during the co-training of our detectors.

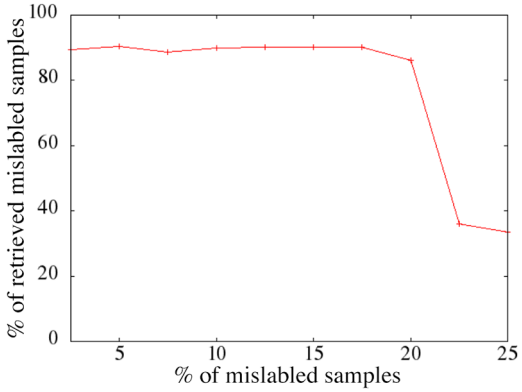


Fig. 7. Impact of the percentage of mislabeled samples on the boosting noise filtering (INRIA dataset).

The impact of the percentage of mislabeled positive samples in the INRIA dataset on the detection performance showed in Figure 9. We can see that the performance degrades slowly from a percentage of positive samples of 0%–10%. The performance starts degrading considerably from 10% and upwards. Note that, no more than 5% of mislabeled samples (non-centered persons, or real mislabeled samples) were noticed in the new generated training samples.

5.2 Collaborative training

We tested the ability of our approach to improve the detectors by comparing the performance of the detectors after 3 iterations (Figs. 11 and 12).

For each iteration (n iterations in Fig. 4), 743 new co-training pairs of images were used (Fig. 5a). For iteration 1, we used pairs from the CTAVIS-11 dataset, for iteration 2, pairs from the CTAVIS-21 dataset and for iteration 3, pairs from the CTAVIS-31 dataset. The AVIS1 testing dataset (used for testing the performance) contains 316 other synchronized pairs of images (similar in resolution to the co-training images shown in Fig. 10). The AVIS dataset contains more challenging synchronized cases than the well known OTCBVS dataset. The initial training of the visible detector has been done using the ATV1 dataset, and the initial training of the infrared detector has been done using the ATI1 dataset. It was noticed that each iteration generated about 300 pairs of new training samples (Fig. 13).

Fig. 8. Examples of visible and far infrared cases filtered by the boosting filter algorithm.

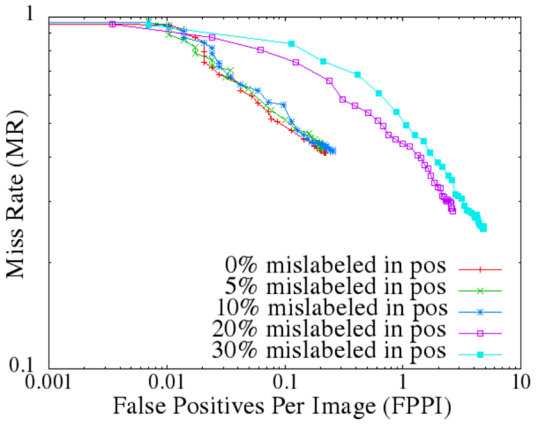


Fig. 9. Impact of the percentage of mislabeled samples on the performance of the ACF detector (trained with INRIA dataset).

Fig. 10. Examples of pairs of visible and infrared images used for the co-training.

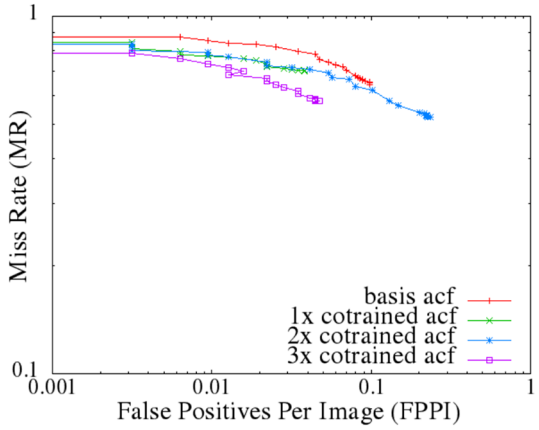


Fig. 11. Detection performance of the ACF detector after 1, 2 and 3 co-training iterations.

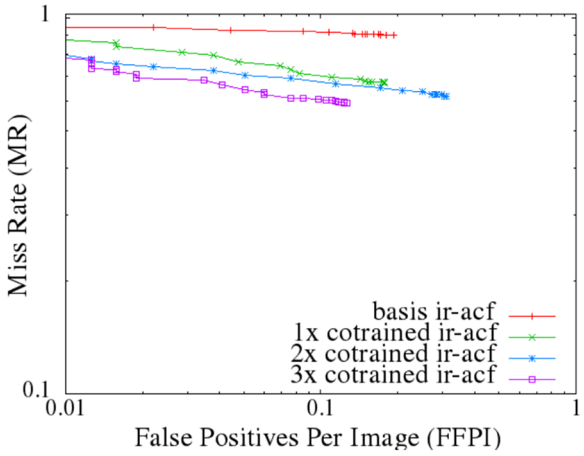


Fig. 12. Detection performance of the IR-ACF detector after 1, 2 and 3 co-training iterations.

Fig. 13. Examples of coupled training data generated using our co-training approach.

As can be seen in Figures 11 and 12, the performance are improved after each iteration. This co-training procedure is an interesting alternative to fully supervised approaches.

6 Conclusion

In this paper, a collaborative training approach for improving human detection was proposed. We showed that the training of a far infrared and a visible human detector can be mutually improved in a semi-supervised manner with the help of a two step procedure which consists of the computation of a thermal objectness score and a step of boosted noise filtering. Our co-training approach is a convenient solution to generate infrared training data that are often not publicly available.

References

1. N. Dalal, B. Triggs, Histograms of Oriented Gradients for Human Detection, in: *Conference on Computer Vision and Pattern Recognition*, 2005
2. P. Dollàr, Z. Tu, P. Perona, S. Belongie, Integral channel features, in: *Proceedings of the British Machine Vision Conference*, 2009, pp. 91.1–91.11
3. P. Dollaár, R. Appel, S. Belongie, P. Perona, Fast feature pyramids for object detection, in: *Transactions on pattern analysis and machine intelligence (TPAMI)*, 2014
4. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: *Advances in Neural Information Processing Systems 28 (NIPS)*, 2015
5. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, *Computer Vision and Pattern Recognition*, 2016

6. L. Zhang, B. Wu, R. Nevatia, I. Systems, L. Angeles, Pedestrian Detection in Infrared Images based on Local Shape Features, in: *Computer Vision and Pattern Recognition (CVPR)*, 2007
7. D. Olmeda, J.M. Armingol, Contrast Invariant Features for Human Detection in Far Infrared Images, in: *Intelligent Vehicles Symposium (IV)*, 2012
8. D. Olmeda, Pedestrian detection in far infrared images, Ph.D. dissertation, 2013
9. R. Brehar, C. Vancea, S. Nedevschi, Pedestrian detection in infrared images using aggregated channel features, in: *Intelligent Computer Communication and Processing (ICCP)*, 2014
10. A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: *COmputational Learning Theory (COLT)*, 1998
11. A. Levin, P. Viola, O.M. Way, Y. Freund, Unsupervised improvement of visual detectors using co-training, in: *International Conference on Computer Vision (ICCV)*, 2003
12. P.M. Roth, C. Leistner, A. Berger, H. Bischof, Multiple instance learning from multiple cameras, in: *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2010
13. S.J. Krotosky, M.M. Trivedi, Mutual information based registration of multimodal stereo videos for person tracking, in: Computer Vision and Image Understanding (CVIU), 2007
14. J.-Y. Bouguet, Jean-yves bouguet's matlab toolbox for calibrating cameras, 2015, http://www.vision.caltech.edu/bouguetj/calib-doc/
15. Z. Shi, T. Wei, M.K. Taghi, Boosted Noise Filters for Identifying Mislabeled Data, *Department of Computer Science and Engineering Florida Atlantic University,* 2005
16. C.L. Zitnick, P. Dollàr, Edge boxes: locating object proposals from edges, in: *European Conference on Computer Vision (ECCV)*, 2013

Cite this article as: Paul Blondel, Alex Potelle, Claude Pégard, Rogelio Lozano, Collaborative training of far infrared and visible models for human detection, Int. J. Simul. Multidisci. Des. Optim. **10**, A15 (2019)