HISTOGRAM OF CONFIDENCES FOR PERSON DETECTION

Lee Middleton, James R. Snowdon

IT Innovation Centre, University of Southampton,United Kingdom
ljm@it-innovation.soton.ac.uk, snowdonjames@googlemail.com

ABSTRACT

This paper focuses on the problem of person detection in harsh industrial environments. Different image regions often have different requirements for the person to be detected. Additionally, as the environment can change on a frame to frame basis even previously detected people can fail to be found. In our work we adapt a previously trained classifier to improve its performance in the industrial environment. The classifier output is initially used an image descriptor. Structure from the descriptor history is learned using semi-supervised learning to boost overall performance. In comparison with two state of the art person detectors we see gains of 10%. Our approach is generally applicable to pretrained classifiers which can then be specialised for a specific scene.

***Index Terms*—** Image analysis, Image classification, Object detection, Identification of persons, Image segmentation

1. INTRODUCTION

Detection of people in images has a long history [1, 2] but despite this has yet to yield good results particularly in cluttered or complex environments. Such environments are prevalent in industry and serve as the motivation for this work. Typical industrial environments are harsh for image processing. They suffer from rapid lighting changes (machinery in operation), occlusion (obscured by equipment), and camera shake (transport of heavy machinery). The environments are also lit to enable the employees to perform their tasks rather than capture them. In our work we have recordings from within such an environment and are examining the problem of person detection. Our resulting method needs to be robust and able to be adaptive. As a starting point for our analysis we examined the most popular person detectors [3, 4].

Whilst the details of the specific approaches differ they do share in common a global threshold to find the final candidates for person location. A specific problem caused by this is that there may be candidate people which are lost by particular thresholds (see figure 1). In this figure four candidate people are found two of which are shop window manikins. However, notice that the manikins have higher confidence (0.39 and 0.29) than one of the people (0.13). In this case if the global threshold was adjusted to remove the false positives then one human would also disappear. This means that there needs to be an alternate way of finding the threshold in the image. Ideally, each pixel should have a different threshold value.

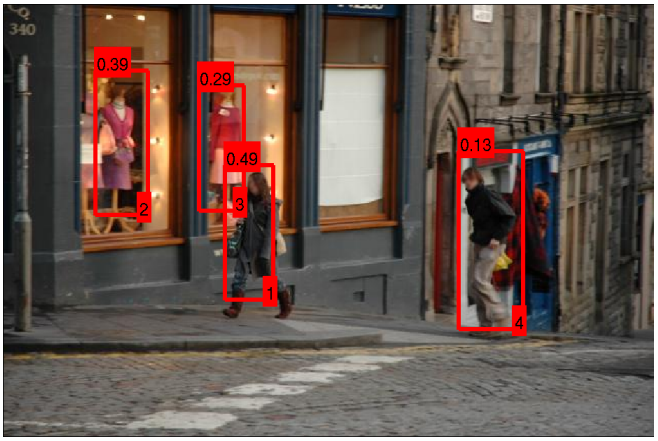
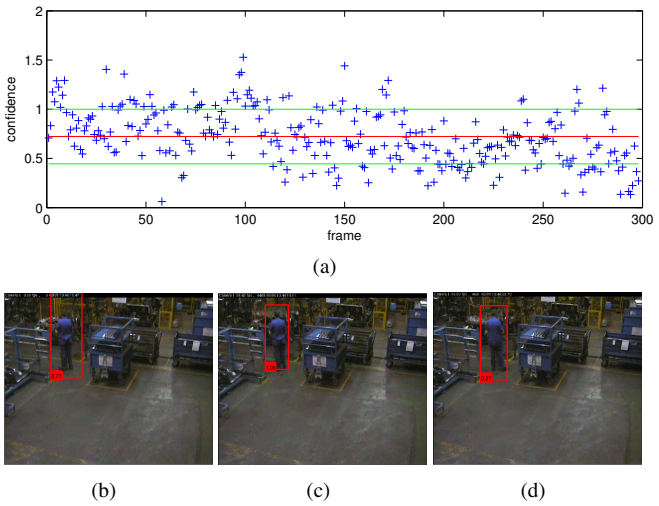


Fig. 1. Confidences and ranking for a person detector.

In addition to needing a spatially localised threshold there is a second problem that can occur. For a single stationary individual the confidence values are not constant. This is illustrated in figure 2. Notice there is large variation of the confidence from frame to frame. The mean value (as shown by the red line) is 0.7 but the standard deviation is comparatively large at 0.3 (green lines). Such a variation makes it difficult to assign a single threshold even to small regions of the image. Additionally, low thresholds will result in many false positives. For these reasons it is a requirement to vary the threshold as the sequence progresses. It should be evident that smoothing approaches are unlikely to work here as the difference between consecutive frames can be very large.

The failure of global thresholding due to lack of spatial and temporal support will be addressed in the next sections. Specifically, we will attempt to improve the overall response of the base person detector by adapting the results of the existing classifier using semi-supervised learning. Within this we make the assumption that there are confidence values for each frame and each frame region which are good indicators

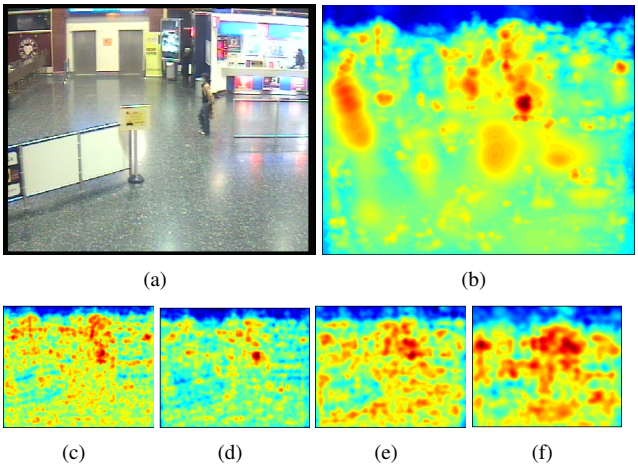
This work is supported by the EC Framework 7 Program (FP7/2007-2013) under grant agreement number 216465 (ICT project SCOVIS).



(a)

(b) (c) (d)

Fig. 2. (a) Variation of confidence values for a stationary individual for 300 frames (b) frame 1 (confidence=0.71) (c) frame 150 (confidence=1.08) (d) frame 300 (confidence=0.27)



(a) (b)

(c) (d) (e) (f)

Fig. 3. Confidence maps found from person detection results (a) Original image (b) Final confidence map (c-f) Confidence maps at different scales.

of there being a person. We will discover these by learning from the history of the classifier confidence values.

2. APPROACH

We will start with outlining how a typical person detector works. For a given image I, which is a single frame in sequence of images with infinite past and future we can apply a person detector, H. The output, P, from the person detector is a number of bounding boxes with associated confidences. These can be ordered by the region and the scale they correspond to. P has 59000 results for [3] and 47000 for [4]. Typically, a global threshold is applied to P to find the best candidates for people, P_t. Rather than using P_t we propose to use the original results from the person detector P. The motivation here is to exploit relationships from within the data and history to improve the performance. Thus we have an exhaustive list of boxes for I:

$$P = \{w_{s,i} : s = [1, S], i = [1, N_s]\}$$

Here S is the number of scales in the person detector (our method works equally for $S = 1$), and N_s is the number of windows at that particular scale. The window, $w_{s,i}$, is defined by the location, size, and confidence. It is a vector, $w_{s,i} = [x, y, w, h, c]$. Without loss of generality we assume that the confidence can be considered to act at the centre of the box. This is a fair assumption as the best match for H will occur when the person fills the entire box and is centred on it. We also analyse each scale independently. As an example, a person is typically valid for a number of detections across different scales. Thus we can treat all scales as indepedent to find the best detection scale. The classifier output at a single scale is a sampling of the original data on a uniform lattice. Consequently we can interpolate over this data to produce an image of the confidences at a single scale. Figure 3 is an example of the approach applied to an example image. The images along the bottom are confidence maps created at a single scale from P. The final confidence map, C, can be created from the individual confidence map images, C_s, by a maximisation process. This is similar in spirit to the approach of [5]. This novel generation of a confidence map from bounding box information underpins the rest of our approach. Additionally, it is broadly applicable to other region oriented detectors.

The confidence map could be considered to be an image descriptor. In our work we plan to learn from the time history of the confidences. This is unusual but has an intuitive basis. The output from a person detector can be considered a statistical sampling of the space of detected people in the image. By integrating history we can improve the proposed distribution. Statistical classification based on this distribution can provide us with better classifications of the person. Figure 4 illustrates two examples of the history of the confidences with a window size of 700 frames. Contrasting the two cases in the figure shows there to be clear differences between them. Firstly, in the case where there is a person the confidence values are much higher. Secondly, the addition of the person creates a second distribution centred on the person as seen by the small peak centred about 0 in figure 4(a). Thus, the introduction of a person to a scene will result in a slow shift of the distribution to positive numbers. Whereas, a scene without a person will yield a distribution generally centred on low negative numbers. Using these observations we can model the resulting time-series histograms using mixture models. This approach is inspired by [6].

Our mixture models were made up of N_G Gaussian dis-

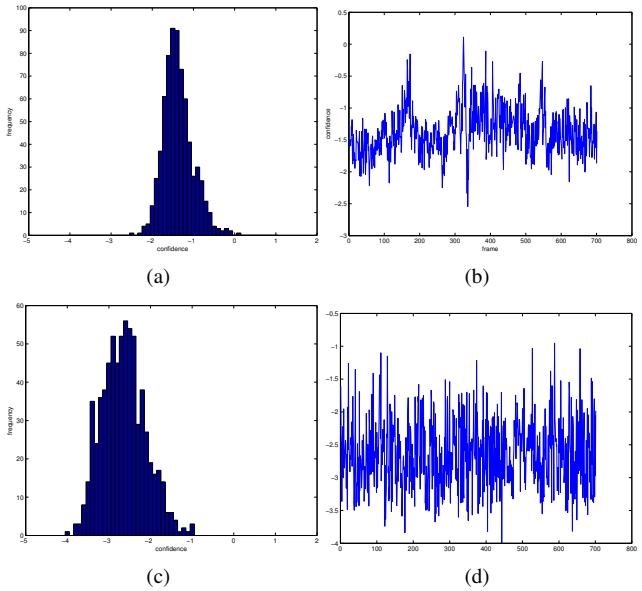


Fig. 4. Time-series of the confidence values for 700 frames (a) histogram for person (b) raw confidences for a person (c) histogram for no person (d) raw confidences for no person

tributions. Each distribution has an associated weight $\alpha_{k,t}$, mean $\mu_{k,t}$, and standard deviation $\sigma_{k,t}$ for distribution k and frame t. As every pixel in the image has a different history there is a mixture model for every pixel. Updating the mixture models occurs using a simplified form of expectation maximisation as outlined in [6]. We start with the current confidence value for a pixel, x. If this fails to match any of the existing Gaussians then we replace the most unlikely distribution with a new one with mean the same as the confidence value, a large variance, and a low weight. For a specific frame, t, the weights are updated:

$$\alpha_{k,t} = \begin{cases} (1-\lambda)\mu_{k,t-1} + \lambda x & \text{best match} \\ (1-\lambda)\mu_{k,t-1} & \text{other cases} \end{cases}$$

Here λ is the learning rate and controls how quickly the distributions incorporate the new data. As this rule employs an exponential window it is related to the size of the history. In the case that x lies within $2.5\sigma_{k,t}$ (98%) of the mean then the mean and variance of the best match are updated to integrate the new data:

$$\mu_{k,t} = (1-\rho)\mu_{k,t-1} + \rho x$$
$$\sigma^2_{k,t} = (1-\rho)\sigma^2_{k,t-1} + \rho(x-\mu_{k,t})^T(x-\mu_{k,t})$$

Where ρ is learning rate multiplied by the Gaussian for this confidence value. This process is carried out for every point in the image with a number of Gaussians modelling the behaviour of the confidence history. Labels are assigned to each of the Gaussians depending on the observed properties of them. Typically lower means are considered to be background distributions whereas higher ones are considered to belong to people. Labels are changed if the location of the distribution drifts significantly or the ratio of $\frac{\alpha_{k,t}}{\sigma_{k,t}}$ changes to resemble any of the other classes. The process of assigning labels based on learned classifier results is typical of semi-supervised learning approaches.

Taking the confidence map and generative models together we can build a system to improve the person detection. It initially proceeds by applying a person detector to the data. Before on-line operation a suitable history is needed to create the histograms of confidences and to initialise the Gaussian models for each image pixel. Practically, we find that this is no more than twice the window size. Once the initial labelling is performed this system can be used for classification. This proceeds by first using the Gaussian models to perform a classification at each pixel. We have three classes for this person, background, and uncertain. The classes are decided upon the response of the current confidence value and to which Gaussian they belong. Results that are classified as person and background are then used to augment the existing Gaussian models using the update rules previously outlined. As a final step the resulting classification undergoes a morphology step to reject noise. We then perform a non-maximal suppression step to find the best candidate for the person. In essence our approach could be considered to be using semi-supervised learning [7] to adapt the classifier output.

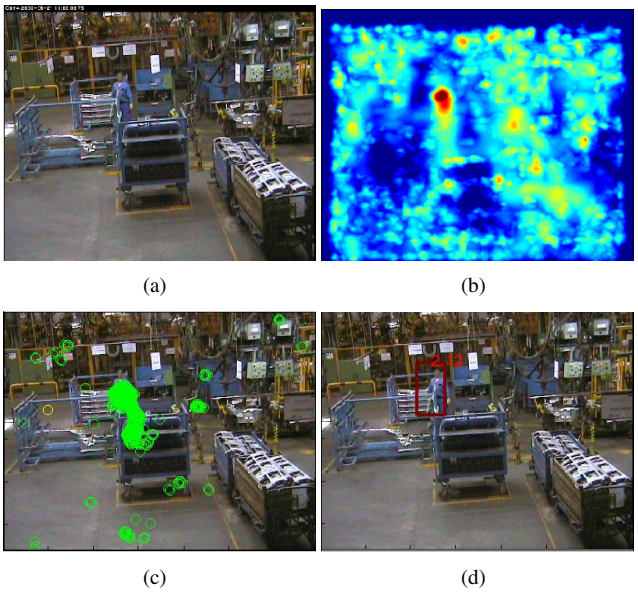


Fig. 5. Representative results (a) Original image (b) confidence map (c) classification (d) resulting bounding box

(a) (b) (c)

Fig. 6. Bounding box detection (a-b) industrial (c) i-Lids

3. EXPERIMENTS

This section presents preliminary results from the system applied to several different data sets. The data sets we chose were two separate collections (6 months apart) from a large manufacturing environment and the i-Lids data set. We chose one subset of each for experimentation taking approximately 1400 frames for each set. A window size (N_w) for the time-series histograms of 200 frames was chosen and a further 200 frames were used to initialise the Gaussian models. Figure 5 shows an example image with the various important steps in our system. Notice that our adaptive technique reduces the search space for people dramatically. Specifically, results are concentrated around the region where historically a person has been. The false positives in the image are mostly removed by the application of morphology. Some more examples of the result on the data sets examined are shown in figure 6. The three examples show the systems ability to find people in occluded environments. Furthermore, in the case of i-Lids it finds two people at very different scales.

For a better comparison of performance we examined the result of our person detector versus the ones from [3] and [4]. To do this we marked up ground truth for the entire of the sequences outlined previously. This gave us 2479 frames worth of data with 3200 bounding boxes. Then we obtained the number of false positives (F_p) and true positives (T_p) in the entire sequence in each of the three cases with the same fixed threshold (rejects 90% of boxes). From the ground truth data we know the number of real positives, P, for the data. Then we computed the precision ($P_r = \frac{T_p}{T_p+F_p}$), recall ($R = \frac{T_p}{P}$), and F-measure ($F_m = \frac{2P_rR}{P_r+R}$) to compare the aggregate results. These are illustrated in table 1. Our approach used the [3] person detector as a source of the confidence information. The results show our proposed approach having an edge over the other approaches.

	P_r	R	F_m
Felzenszwalb et al. [3]	0.85	0.65	0.74
Dalal & Triggs [4]	0.73	0.56	0.63
Our Approach	0.93	0.72	0.81

Table 1. Comparison of the three approaches on dataset

4. CONCLUSION

In this paper we presented a novel approach for improving person detection based on using the output of an existing classifier as an image descriptor. Salient features from the history of this descriptor are learned via semi-supervised learning to improve the classification task. We presented several main contributions. Firstly, we reinterpreted the bounding box data as a confidence map which can be examined on a per pixel fashion. Secondly, we proposed that useful information could be learned from the history of these confidences. Finally, we presented a mixture method to model this descriptor and learn in an on-line fashion a correction which gives a better classification of people. Currently, the work is ongoing but is showing promising results. We are currently looking at the efficacy of the descriptor. Additionally we would like to apply our technique to correct other pretrained classifiers. Lastly, we plan to speed the performance of the approach by looking at simpler methods to model the descriptor behaviour.

5. REFERENCES

[1] K Akita, "Image Sequence Analysis of Real World Human Motion," *Pattern Recognition*, vol. 17, no. 1, pp. 73–83, 1984.

[2] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition*. June 2009, pp. 304–311, IEEE.

[3] P. F. Felzenszwalb, D. McAllester, and D. Ramanan, "A Discriminatively Trained, Multiscale, Deformable Part Model," in *Proceedings of the IEEE Internal Conference on Computer Vision and Pattern Recognition*. 2008, IEEE.

[4] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2005, pp. 886–893, IEEE.

[5] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 3, pp. 411–26, March 2007.

[6] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 246–252, 1999.

[7] X. Zhu, "Semi-supervised learning literature survey," Tech. Rep. 1530, Computer Sciences, University of Wisconsin, 2005.