Robust Human Detection under Occlusion by Integrating Face and Person Detectors

William Robson Schwartz[1], Raghuraman Gopalan[2], Rama Chellappa[2], and Larry S. Davis[1]

[1] University of Maryland, Department of Computer Science
College Park, MD, USA, 20742
[2] University of Maryland, Department of Electrical and Computer Engineering
College Park, MD, USA, 20742

Abstract. Human detection under occlusion is a challenging problem in computer vision. We address this problem through a framework which integrates face detection and person detection. We first investigate how the response of a face detector is correlated with the response of a person detector. From these observations, we formulate hypotheses that capture the intuitive feedback between the responses of face and person detectors and use it to verify if the individual detectors' outputs are true or false. We illustrate the performance of our integration framework on challenging images that have considerable amount of occlusion, and demonstrate its advantages over individual face and person detectors.

1 Introduction

Human detection (face and the whole body) in still images is of high interest in computer vision. However, it is a challenging problem due to the presence of variations in people's poses, lighting conditions, inter- and intra- person occlusion, amongst others. Occlusion, in particular, poses a significant challenge due to the large amount of variations it implies on the appearance of the visible parts of a person.

There are many human detection algorithms in the literature. In general, they fall into two categories: subwindow-based and part-based approaches. In the former category, features extracted from subwindows located within a detection window are used to describe the whole body. Subwindow-based approaches can be based on different types and combinations of features, such as histograms of oriented gradients (HOG) [1], covariance matrices [2], combination of several features [3], and multi-level versions of HOG [4]. On the other hand, part-based approaches split the body into several parts that are detected separately and, finally, the results are combined. For instance, Wu and Nevatia [5] use edgelet features and learn nested cascade detectors for each body part. Mikolajczyk et al. [6] divide the human body into seven parts, and for each part a cascade of detectors is applied. Shet and Davis [7] apply logical reasoning to exploit contextual information augmented by the output of low level detectors.

M. Tistarelli and M.S. Nixon (Eds.): ICB 2009, LNCS 5558, pp. 970–979, 2009.

Fig. 1. Image where occlusion is present and fusion of detectors can increase detection accuracy. Face of person b is occluded. Once the legs and torso are visible, results from a part-based person detector can be used to support that a human is present at that location. On the other hand, the legs of person c are occluded, in such a case, face detector results can be used to reason that there is a person at that particular location since the face of person c is perfectly visible.

Subwindow-based person detectors present degraded performance when parts of the body are occluded; part-based approaches, on the other hand, are better suited to handle such situations because they still detect the un-occluded parts. However, since part-based detectors are less specific than whole body detectors, they are less reliable and usually generate large numbers of false positives. Therefore, to obtain more accurate results it is important to aggregate information obtained from different sources with a part-based detector. For this, we incorporate a face detector.

Face detection is an extensively studied problem, and the survey paper [8] provides a comprehensive description of various approaches to this problem. For example, Viola and Jones [9] use large training exemplar databases of faces and non-faces, extract feature representations from them, and then use boosting techniques to classify regions as face or non-face. Other algorithms, for instance Rowley et al. [10], uses a neural network to learn how the appearance of faces differ from non-faces using training exemplars, and then detect faces by seeing how well the test data fits the learned model. Another class of approaches, exemplified by Heisele et al. [11], uses a part-based framework by looking for prominent facial components (eyes, nose etc), and then uses their spatial relationship to detect faces. Although such methods are more robust to image deformations and occlusions when compared with holistic approaches, the choice of feature representations and accurate characterization of the relationships between the facial components is still a challenge.

The question that arises naturally is then, how to fuse these two sources to improve overall detection performance. Specifically, is it possible to use the response profiles of the two separate detectors, to reinforce each other, as well as provide a basis to resolve conflicts? This is the question we address in our work.

Figure 1 motivates the utility of combining face and person detectors. First, while the lower half of person c is occluded, the face detector can still detect the face of the person, whereas the person detector might fail. Nevertheless, we can try to *explain* the response of the person detector based on the response of the face detector, and conclude that a person is present. Another case is the reverse situation such as b and d in Figure 1 whose faces are partially occluded while the body parts are completely visible. Such situations occur often in real-world scenarios, and motivates exploring *feedback* between face and people detectors.

2 Face and Person Detection

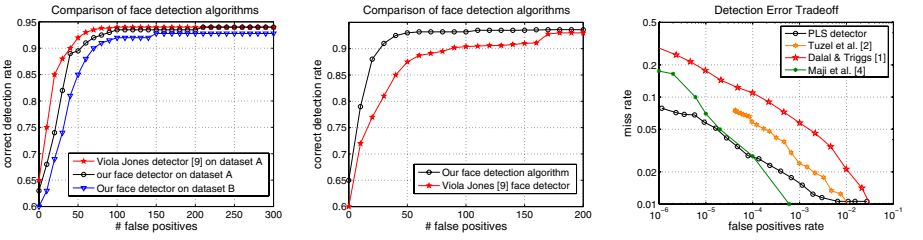
In this section we give a synopsis of our algorithms for face detection and person detection. We also provide detection results of applying the individual algorithms on standard datasets, showing that these detectors individually achieve results comparable to state-of-art methods. However, a point to keep in mind is that these standardized datasets do not have considerable amounts of occlusion, which is the main problem that we address in our work.

2.1 Face Detection

We use a feature-based approach to detect faces from still images. Our approach, motivated by [12], is based on using an optimal step edge operator to detect shapes (here, the facial contours are modeled as ellipses). The crux of the algorithm is then to obtain the edge map of the image using a derivative of double exponential (DODE) operator, and fit various sized ellipses to the edge map. Image regions that have high response to ellipse fitting signify locations that likely contain faces.

We then conduct post-processing on these short-listed regions by computing three different cues - color [13], histogram of oriented gradients [1], and eigenfaces [14], and combine the three feature channels using support vector machines [15] to decide whether a face is present or not. The motivation behind the choice of these descriptors is: (i) the human face has a distinct color pattern which can be characterized by fitting Gaussian models for the color pattern of face regions, and non-face regions; (ii) the histogram of oriented gradients capture the high interest areas in faces that are rich in gradient information (eyes, nose and mouth) that are quite robust to pose variations, and (iii) eigenfaces captures the holistic appearance of the human face. These three feature channels capture a mix of global and local information about the face, and are robust to variations in pose.

Our algorithm was tested on the MIT+CMU face dataset [10]. This dataset has two parts. The first part (A) has 130 frontal face images with 507 labeled faces, the second part (B) has 208 images containing 441 faces of both frontal and profile views. The results of our algorithm are presented in Figure 2(a). Most other algorithms that are evaluated on this dataset do not provide the full ROC, but rather provide certain points on the ROC. Since Viola and Jones [9]



(a) MIT+CMU dataset (b) Maritime dataset (c) INRIA person database

Fig. 2. Experimental results for face and person detection

quote their ROC for part A of this dataset, we have compared our ROC with theirs; even otherwise, it can be observed that our performance is comparable to the ROC points of other algorithms (like Rowley et al. [10]). Since we are interested in detecting partially occluded faces we also compare our approach to the OpenCV implementation of Viola and Jones [9] method on the internally collected maritime dataset in Figure 2(b).

2.2 Person Detection

For person detection we use a method that combines HOG [1] and features extracted from co-occurrence matrices [16]. For each detection window, features extracted from HOG and co-occurrence matrices are concatenated and projected onto a set of latent vectors estimated by the partial least squares (PLS) method [17] in order to reduce the dimensionality of the input vector. The vector obtained after dimensionality reduction is used as the feature vector to describe the current detection window. Finally, the feature vector is classified by a quadratic classifier as either human or non-human sample. As a result, we obtain a probability estimate. Figure 2(c) shows comparisons using the INRIA person dataset [1]. Like face detection, the person detection approach used also achieves results comparable to state-of-art person detectors [1,2,4].

Since part-based approaches are better suited to handle situations of occlusion, we split the person detector into seven different detectors, which consider the following combinations of regions of the body: (1) top, (2) top-torso, (3) top-legs, (4) torso, (5) torso-legs, (6) legs, and (7) full body, as illustrated in Figure 3. Therefore, at each position in the image the person detector estimates a set of seven probabilities. The training for these detectors was performed using the training set of the INRIA person dataset.

As discussed in the literature survey, part-based approaches for person detection have been employed previously. Here, we use a part-based approach in tandem with a face detector creating a small number of intuitive case-based models for overall person detection.

Although the face and person detectors present results comparable to the state of the art on these datasets, these algorithms face difficulties when there is significant occlusion. To this end, we explore how to overcome this problem by combining the responses of the individual detectors.

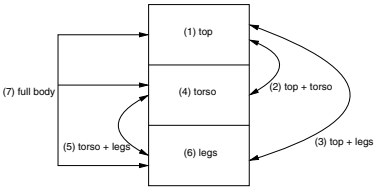


Fig. 3. Parts of a detection window used to train multiple detectors

3 Integrating Face and Person Detection

In this section we present our algorithm for integrating the response profiles of face and person detectors. We model observations of the individual detectors, and generate hypotheses that capture intuitive relationships between the responses of the face detector and the person detector. Specifically, we describe a set of situations where the output of one detector can be logically combined with the other detector's output to eliminate false alarms or confirm true positives.

3.1 Modeling the Response Profiles of the Individual Detectors

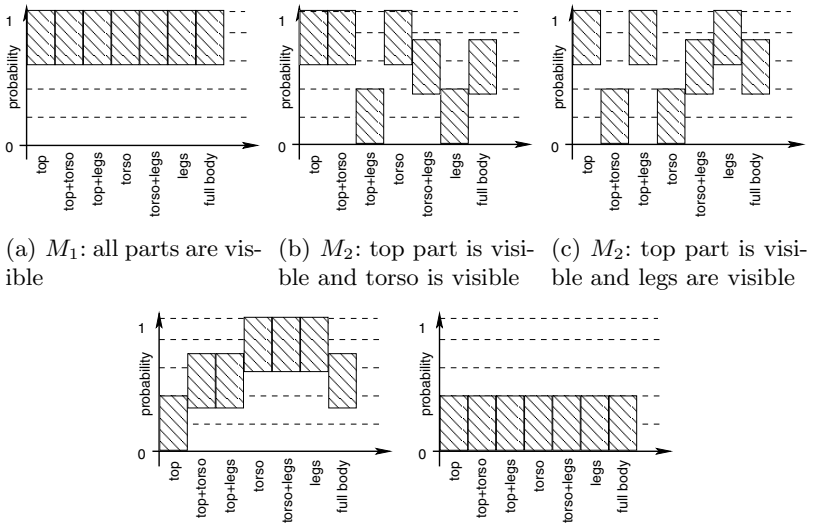
To integrate person and face detectors' output we first create models according to the probability profile resulting from individual detectors (the seven probabilities from part-composition person detector and one from the face detector).

For the person detector, we summarize the probability profile obtained by the seven probabilities into a set of four models that inherently capture situations in which various combinations of face and person parts are detected with *high* probability. Specifically,

Model M_1: all body parts are visible
Model M_2: top is visible, torso and legs may or may not be visible. This corresponds to the typical situation in which a person's legs are occluded by some fixed structure like a desk, or the railing of a ship.
Model M_3: top is invisible, whereas torso and legs are visible
Model M_4: all body parts are invisible

Given the set of seven probabilities estimated by the person part-combination detectors, we define probability intervals that characterize each model. The estimation of the intervals for models M_1 and M_4 can be done automatically by evaluating probability of training samples from standard person datasets. However, probability intervals for models M_2 and M_3 only can be estimated if a training set containing partially occluded people were available. Due to the absence of such dataset, we define the probability intervals for M_2 and M_3 manually.

Figure 4 shows the probability intervals for each model. A model M_i fits a detection window if all seven estimated probabilities fall inside the probability intervals defined by M_i. We also estimate a degree of fit of a detection window to each model by simply counting the number of probability intervals satisfied by the response profile:



(a) M_1: all parts are visible (b) M_2: top part is visible and torso is visible (c) M_2: top part is visible and legs are visible

(d) M_3: top part is invisible (e) M_4: all parts are invisible

Fig. 4. Models designed considering the output profile of the person detector. The x-axis has the seven detectors and the y-axis the probability interval for each one according to the model. Note that M_2 has two sub-cases, shown in (b) and (c).

$$f(M_i) = \sum_{j=1}^{7} \begin{cases} 1 & \text{if } u_{i,j} \leq P_j \leq l_{i,j} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where P_j denotes the probability estimated by the j-th detector, $u_{i,j}$ is the upper bound for the j-th interval defined for M_i and $l_{i,j}$ denotes the lower bound. Therefore, we can rank the models according to how well they fit a given detection window. We say that a model M_i has a rank higher than M_j when $f(M_i) > f(M_j)$.

For the face detector, the observations are characterized by the probability values indicating the presence of face for a given detection window. According to this probability we define three models. We say that a face is present if the probability exceeds a certain threshold (model F_1). We also consider the case when the probability is smaller than the threshold but not negligible (i.e. face might be partially occluded), we refer to this as model F_2. Model F_2 is interesting when the person detector gives a response that supports the low (but not negligible) confidence of the face detector. Finally, we say that a sample fits model F_3 if the probability of face detector is very low.

3.2 Generating Hypotheses to Integrate Detectors

Now that we have designed models according to the response profiles to capture occlusion situations, we create a set of hypotheses (rules) to characterize the

relation between the detector responses so that these different sources of information can be used to *verify* each other's output. We separate the possibilities into five different hypotheses. The first two hypotheses describe the scenario where the person detector (PD) is used to *verify* the output of the face detector (FD), and the remaining three hypotheses deal with the alternate scenario of using face detector to verify the person detector outputs. The hypotheses are described in the form of conditional rules as follows.

$H_1 : [(f(M_1) \wedge f(M_2)) > (f(M_3) \wedge f(M_4))|F_1]$ Given that the face detector provides high response for a detection window, we look at the models that characterize the person detector output. Since the face is visible, the output of PD should better fit models M_1 or M_2 than M_3 and M_4 since we expect the top (head and shoulder) features to be detected by the person detector. If that is the case, then PD output verifies that the FD output is correct. Thus, a person is present at that location.

$H_2 : [(f(M_3) \vee f(M_4)) > (f(M_1) \wedge f(M_2))|F_1]$ The alternate case is, given high response for the face detector, if the output of PD fits either M_3 or M_4, then PD indicates that the face is not visible, and hence the output of the FD is a false alarm.

$H_3 : [(F_1|(f(M_1) \vee f(M_2)) > (f(M_3) \wedge f(M_4))]$ Given that the rank of M_1 or M_2 is greater than M_3, if FD gives a high response, then the face detector is reinforcing the output of the person detector. Thus, we conclude that a person is present at the corresponding location.

$H_4 : [(F_2|f(M_3) > (f(M_1) \wedge f(M_2) \wedge f(M_4))]$ A slightly different case from H_3 is when FD has low response, but still has some probability higher than 0 but not high enough to conclude the presence of face. In this case, if for the person detector the rank of M_3 is higher than M_1, M_2, and M_4, then we still decide that there is a person whose face is partially occluded. This is because M_3 captures the situation where the face is occluded, while the torso and legs are visible.

$H_5 : [F_3|(f(M_1) \vee f(M_2) \vee f(M_3)) > f(M_4)]$*:* This final hypothesis deals with the case where the output of person detector fits either M_1, M_2, or M_3, and the probability outputted by the face detector is negligible, so that it cannot come under H4. In such a case, since the face is completely invisible, we decide that the PD output is a false alarm.

Essentially, the above hypotheses are built on the fact that the presence of the face implies the presence of a person and vice-versa. We do need some confidence value for the presence of face to make decisions on the output of the person detector. This is based on our observation that the presence of just the torso and legs with no information regarding the face is not a strong cue to detect a person. This condition gives rise to many false alarms.

4 Experimental Results

In this section, we demonstrate with experiments how our integration framework improves detection under occlusion, as well as reduces the false alarms. We tested

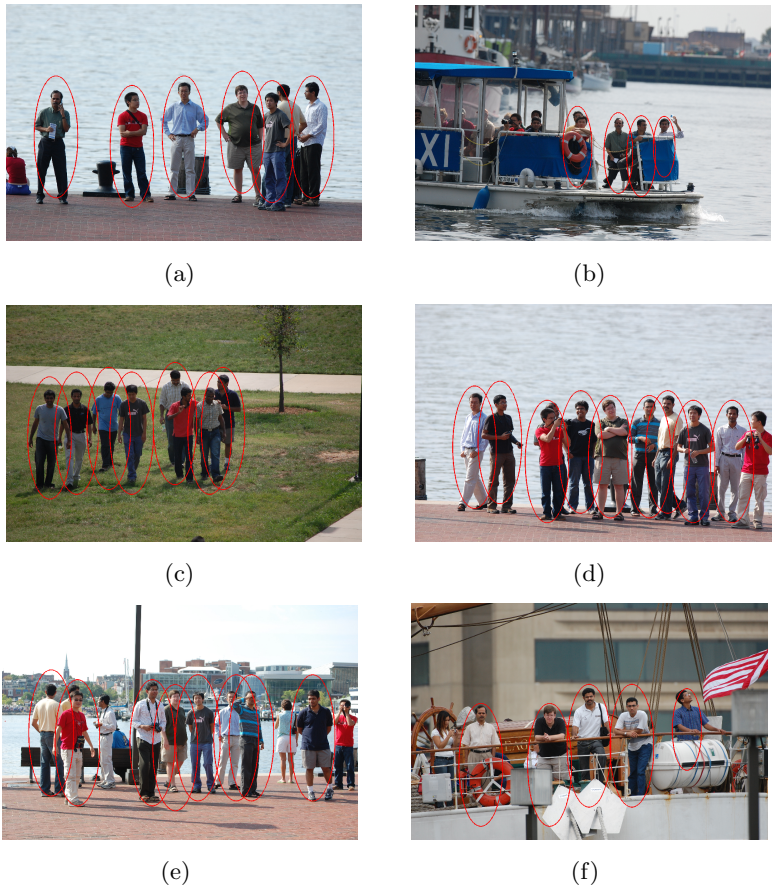


Fig. 5. Results on images from maritime dataset (better visualized in colors)

our algorithm on challenging images taken from an internally collected maritime dataset. It contains images of 3008×2000 pixels, which is suitable for face and person detection, unlike standard datasets used for person detection, which in general contain images with resolution too low to detect faces. This dataset is a good test-bed since it provides challenging conditions wherein the individual face/person detector might fail, thereby emphasizing the need to fuse information obtained by these detectors.

We now present several situations where the integration framework helps to detect humans. In the image shown in Figure 5(b) a person detector would fail to detect people seated since the lower body is occluded. However, our framework combines face information with the presence of the top part of the body (head and shoulders) captured by the person detector. Therefore, it concludes that a person is present. Additionally, Figures 5(c), (e), and (f) contain people who are partially occluded. Such conditions would reduce significantly the probability estimated by an independent person detector, whereas the integration helps resolve this problem.

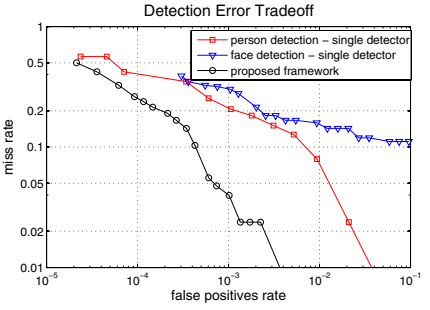


Fig. 6. Detection error tradeoff comparing the integration to individual face and person detectors. The proposed framework outperforms the individual detectors for all points on the curve.

Next, if the face is partially occluded, then the person detector output will belong to model M_3, whereas face detector's output will have some small value that is not very high and not negligible either. In this case, the person detector results can be used to identify the presence of the face. For example, Figures 5(d) and (f) contain people whose faces are occluded. In these cases a face detector would fail to give a high response, but the proposed framework overcomes this problem by aggregating information from body parts.

Essentially, since we are using two separate detectors, if the observations of the person detection and face detection provide conflicting information, then our framework mitigates false positives. A typical example is when hypothesis H_2 is satisfied, which can be used to correct the false alarm of the face detector, and when hypothesis H_5 is satisfied, that helps in reducing the false alarms of the person detector. Additionally, if both individual detectors denote the presence of a person, detection is more reliable than when relying on only one detector.

We tested our algorithm on 20 images containing 126 people. Figure 6 presents the detection error tradeoff of our integration method and compares its results to individual detectors. It can be seen that the use of the proposed method results in a substantial improvement in detection accuracy/false alarm suppression. To generate the curve for the our algorithm, we fix the threshold for the face detector and for the person detector we measure how well each model fits a sample by

$$g(M_i) = \frac{1}{7}\sum_{j=1}^{7}\begin{cases} |P_j - u_{i,j}| & \text{if } P_j > u_{i,j} \\ |P_j - l_{i,j}| & \text{if } P_j < l_{i,j} \\ 0, & \text{otherwise} \end{cases}. \tag{2}$$

With this equation we obtain values of $g(M_i)$ for every sample. Then, varying a threshold value from zero to one we are able to evaluate which hypotheses are satisfied at each step.

5 Conclusions

We have described a framework that combines the observations of face and person detector into different models, and makes decisions based on the hypotheses derived from those models. We then demonstrated our algorithm on several challenging images with considerable occlusion, which illustrates the advantages of exploiting feedback between the response profiles of face and person detectors.

Acknowledgements

This research was partially supported by an ONR MURI Grant N00014-08-10638. W.R. Schwartz acknowledges "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior" (CAPES - Brazil, grant BEX1673/04-1).

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, vol. 1, pp. 886–893 (2005)
2. Tuzel, O., Porikli, F., Meer, P.: Human detection via classification on riemannian manifolds. In: CVPR, pp. 1–8 (2007)
3. Wu, B., Nevatia, R.: Optimizing discrimination-efficiency tradeoff in integrating heterogeneous local features for object detection. In: CVPR, pp. 1–8 (2008)
4. Maji, S., Berg, A., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: CVPR, pp. 1–8 (2008)
5. Wu, B., Nevatia, R.: Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In: ICCV, pp. 90–97 (2005)
6. Mikolajczyk, K., Schmid, C., Zisserman, A.: Human detection based on a probabilistic assembly of robust part detectors. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 69–82. Springer, Heidelberg (2004)
7. Shet, V., Neumann, J., Ramesh, V., Davis, L.: Bilattice-based logical reasoning for human detection. In: CVPR, pp. 1–8 (2007)
8. Yang, M.H., Kriegman, D.J., Ahuja, N.: Detecting Faces in Images: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34–58 (2002)
9. Viola, P., Jones, M.: Robust Real-Time Face Detection. International Journal of Computer Vision 57, 137–154 (2004)
10. Rowley, H., Baluja, S., Kanade, T.: Neural network-based face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 20, 23–38 (1998)
11. Heisele, B., Serre, T., Poggio, T.: A Component-based Framework for Face Detection and Identification. IJCV 74, 167–181 (2007)
12. Moon, H., Chellappa, R., Rosenfeld, A.: Optimal edge-based shape detection. IEEE Transactions on Image Processing 11, 1209–1227 (2002)
13. Hsu, R., Abdel-Mottaleb, M., Jain, A.: Face detection in color images. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 696–706 (2002)
14. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. PAMI, 711–720 (1997)
15. Osuna, E., Freund, R., Girosit, F.: Training support vector machines: an application to face detection. In: CVPR, pp. 130–136 (1997)
16. Haralick, R., Shanmugam, K., Dinstein, I.: Texture features for image classification. IEEE Transactions on Systems, Man, and Cybernetics 3 (1973)
17. Wold, H.: Partial least squares. In: Kotz, S., Johnson, N.L. (eds.) Encyclopedia of Statistical Sciences, pp. 581–591. Wiley, New York (1985)