Haar like and LBP based features for face, head and people detection in video sequences

Etienne Corvee, Francois Bremond

Haar like and LBP based features for face, head and people detection in video sequences

Etienne Corvee and Francois Bremond

INRIA, Pulsar team, Sophia Antipolis, France
{etienne.corvee,francois.bremond}@inria.fr
http://www.inria.fr/equipes/pulsar

Abstract. Actual computer vision algorithms cannot extract semantic information of people activity coming from the large and increasing amount of surveillance cameras installed around the world. Algorithms need to analyse video content at real time frame rate and with a false alarm detection rate as small as possible. Such algorithms can be dedicated and specifically parameterised in certain applications and restrained environment. To make algorithms as useful as possible, they need to tackle many challenging issues in order to correctly analyse human activties. For instance, people are rarely entirely seen in a video because of static (contextual object or they are partly seen by the camera field of view) and dynamic occlusion (e.g. person in front of another). We here present a novel people, head and face detection algorithm using Local Binary Pattern based features and Haar like features which we refer to as couple cell features. An Adaboost training scheme is adopted to train object features. During detection, integral images are used to speed up the process which can reach several frames per second in surveillance videos.

Keywords: face, head, people detection, Local Binary Pattern, Haar

1 Introduction

A large variety of video applications requires objects of interest to be detected, recognized and tracked in a particular scene in order to extract semantic scene information before being treated for activity modelling. In particular, most video surveillance applications rely on the detection of human activities captured by static cameras. In this domain, although cameras remain mostly fixed, many issues occur. For example, outdoor scenes can display varying lighting conditions (e.g. sunny/cloudy illumination, shadows), public spaces can be often crowded (e.g. subways, malls) and images can be obtained with a low resolution and can be highly compressed. Hence, detecting and tracking objects in such complex environment remains a delicate task to perform. Although the techniques presented in the state of the art of this domain show great results, their success is relative to the environment, the camera location as well as the evaluation context in which the techniques are tested. One of the major difficulty is encountered

when detecting and tracking humans in occlusion scenarios since their bodies overlap onto the image plane and their foreground pixels cannot be separated when they are simply thresholded from a background reference frame. Therefore vision algorithms need to extend their analysis using information held by the underlying pixels.

In this paper, attention is focused in detecting face, head and people body in videos. Faces can be seen as small objects when people are distant from the camera. People need to face the camera before the face can be detected. The detection is difficult due to motion blurring and image compression artefacts. Such small objects as faces are difficult to detect thoroughly in the sequences and are hence difficult to track. Even when people face more or less the camera, their movement amplitude can be large compared to the face size. However, heads are bigger objects and can be tracked in all directions of a person moving direction. The only discomfort for head detection is the absence of complex visual signature as head can have many other object visual aspect (e.g. triangles as woman untied long hair, round objects as Jackson five-like hair style). Therefore, head should not be detected without constraints. People silhouette has a more complex and diverse visual signature. People are much larger objects than faces and heads and thus they are easier to track when they are walking. The main drawback with people detection is caused by occlusion and people tracking should always be combined with body parts such as head and face for people to be efficiently tracked and recognised.

Face detection has been studied for many decades with a much higher interest this last decade since face detection and recognition algorithms are getting performant in terms of processing cost and can be used for a large variety of applications such as in security or multimedia applications. The algorithm developed by Viola and Jones [23] and distributed by the OpenCv library [16] is widely used for face detection. They use Haar features to represent faces and an Adaboost [6] algorithm to build a cascade of fast classifiers. Haar features take advantage of the grey level differences between regions of the face. For example, Suguna *et al* [19] first quickly extract face candidates using the most significant face feature i.e. the eyes, and then use a SVM classifier on histogram equalised candidates to determine if they really are faces or not. Many techniques exist as reviewed in the survey of Yang *et al* [26] with performances fairly comparable depending on the applications and their use. For example, McKenna *et al* [12] filter out false detections and increase the speed of their algorithm by detecting objects of interest in moving image regions. They use a Gabor Wavelet Transform to model features of faces and extract the eigen poses of faces captured under different rotation angles and many different lighting orientation with respect to the camera position. Many face tracking algorithms take advantage of fast colour segmentation of the skin to build a fast face tracking algorithm. However, the majority of these techniques deal with only one class of skin colour [17, 9, 22].

Another important features used for object detection is provided by the calculation of Histogram of Oriented Gradients i.e. HOG. Pedestrians, faces and bicycles are successfully detected when represented by HOG [4, 1]. Such as with

Haar based detectors, a boosting technique is also often used to model and rapidly detect objects [10] such as humans [27]. HOG features are extracted from selected areas of the image and compared to the trained models for object classification. HOG features can also be tracked independently without having to classify objects [1]. The detection of objects can be constrained with object motion information given, for example by optical flow of pedestrians [4].

Haar features have been well studied for the detection of objects, in particular for face detection [23]. Histograms of Oriented Gradients i.e. HOG have been successively used for object detection [5] such as for pedestrians, faces and bicycles [4, 1]. A boosting technique is often used to model and rapidly detect objects [10] such as humans [27]. SVM coupled with HOG is used in [4] for this task and successively implemented for the OpenCv library. Although Covariance features can be computationally expensive to estimate, they have strong discriminative powers. Tuzel and al. [21] use a Logiboost algorithm on Riemannian manifolds. Covariance features in a Riemannian geometry are trained allowing the classification of pedestrians. More recently, high performances were obtained by [7] using a highly trained set of granules.

Many recent papers use body parts to enhance people detection performance. There are many ways to combine body parts; for instance Mohan et al. [15] studied different voting combination of body parts classifiers. In [14], Mikolajczyk et al. use 7 body part detectors independently trained to better detect humans. Hussein and Porikli [8] introduce the notion of deformable features in a Logiboost algorithm to allow body parts to have non fixed locations in a people image template. Hierarchical trees has shown great interests to classify multiple object classes into clusters as performed by Mikolajczyk et al. [13]. The SIFT feature is more recently used as object feature for people detection. A 128 dimensional SIFT [11] descriptor is used on the dominant edge orientation of feature regions. They have tested their techniques for recognizing pedestrians, cars, motorcycles, bicycles and rocket propelled grenade launcher shooters in a large image database.

In terms of tracking in video surveillance applications, performance is best when the tracking scheme is well adapted and used robustly detected objects. Nevertheless, people can never be all successfully detected in crowded scene viewed by a single camera. Moreover, objects interaction in a scene can be complex and rules need to be understood by a tracker to handle difficult cases such as occlusions. For instance a person is allowed to disappear when entering his/her own car or a person has to re-enter a scene after a certain time after entering a cloakroom. A generic tracking algorithm needs to provide consistent people trajectories before robust semantic information can be extracted from a scene. Trackers often use motion prediction model such as Kalman filtering [anonymous] or scene context information [3] to provide consistent people trajectory. Singh et al. [18] first detect high confidence partial segments of trajectories called tracklets. They first detect 4 body parts using a Bayesian combination of edgelet part detector [25] which makes the people detection robust. Using a delay, tracklets of newly detected people can be merged with previously fragmented tracklets.

They have evaluated with success their algorithm in occlusion scenarios present in the Caviar [2] datasets. The tracker uses a multi hypothesis theme: data association is performed using a combination of probabilities obtained from colour model, motion model and a 3D human height model.

In this paper, we present in the first section the features used for face, head and people detection. The features are trained using Adaboost training algorithm. The second section provided details on object detection in video sequences. The face and people detection algorithms are evaluated in the third. A conclusion on the presented features is given in the last section before acknowledgements.

2 People, head and face feature training

The couple cell response feature is defined in section 2.1 to detect faces and a simplified LBP feature is defined in section 2.2 for head and people detection. We apply the same Adaboost training algorithm on each face, head and people features. The training for people detection is done using 7K positive images (5K from NICTA, 1K from MIT and 1K from INRIA people training dataset) and 50 background negative PAL images. Heads are trained with 1K positive images (cropped head images from INRIA and TUD people training datasets) and 10 background negative images. Faces are trained using the standard CMU face database.

2.1 Face detection using couple cell response features

A couple cell feature consists of two adjacent cells $c = \{1, 2\}$ forming a couple of 2 cells. An illustration is given in right image of figure 1. The couple can be oriented horizontally, vertically or in diagonal. Pixel distributions v_1 and v_2 of the two cells c_1 and c_2 respectively are compared to give the feature response. A cell pixel distribution is represented by an average E (the expectation term) of the pixel grey levels within this cell. The couple cell response CCR, at feature position and dimension $\chi = \{x, y, w, h\}$, is then calculated from the magnitude and sign of the difference Δv between these two mean values:

$$v_c = E_{\mathbf{x}\epsilon\zeta}\left(I(\mathbf{x})\right) \tag{1}$$

$$\Delta v = v_2 - v_1 \tag{2}$$

$$\mathrm{CCR}_\chi = b\ \mathrm{sign}(\Delta v)\ , \quad 2^b < \Delta v < 2^{b+1} \tag{3}$$

where ζ represents the pixels underlying cell c. The CRR feature is of dimension 17 and its features represent varying ranges of pixel distribution differences:

$$(b; \Delta v) = (-8; [-255 : -128[), (-7; [-127 : -64[)...(0; [-1 : 1[), \tag{4}$$

$$(1; [1 : 2[), (2; [2 : 4[)...(8; [128 : 255]) \tag{5}$$

The couple cell feature is calculated across an image at various scales by varying the cell width w and height h as follows:

$$(w, h) = \{(1,1)(1,2)..(1,8), (2,1)..(8,8)\} \tag{6}$$

2.2 People and head detection using simplified Local Binary Pattern

The standard LBP operator [20] extracts feature vector of size 256. We have implemented a simplified version of the LBP operator by reducing its dimensionality to 16. which we will refer to as the SLBP for the rest of the paper. This SLBP is calculated from 4 cells as illustrated in the left image of figure 1. A pixel intensity average $v_{c,w,h}$ is calculated for each of the 4 cells c, $c = [1:4]$ as defined in equation 7 where E represents the expectation (average) function over pixel locations (x, y).

$$v_c = E_{\mathbf{x}\epsilon\zeta_c}\left(I(\mathbf{x})\right) \tag{7}$$

where ζ_c represent the set of pixels within the cell c area. The first, second, third and forth cell are the top left, top right, bottom left and bottom right cell respectively. The SLBP feature, with feature position and dimension $\chi = \{x, y, w, h\}$, is then calculated from 4 mean pixel intensity differences Δv_q as follows:

$$\text{SLBP}_\chi = \sum_{q=0}^{3} s(\Delta v_q) 2^q \qquad s(.) = \begin{cases} 1 \text{ if } . > V \\ 0 \text{ else} \end{cases} \tag{8}$$

$$\Delta v_q = v_i - v_j \ , \quad (q, i, j) = \{(0,1,0), (1,3,2), (2,2,0), (3,3,1)\} \tag{9}$$

The training is performed on a multiple scale approach by varying the SLBP cell dimensions enumerated in equation 6.

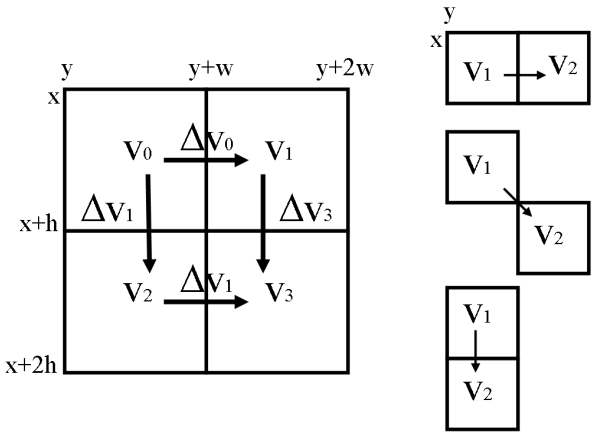

Fig. 1. Left: simplified LBP operator and right: couple cell response feature

3 Adaboost training

For each feature position $\mathbf{x}$ and dimension width w and height h that we encapsulate into $\chi = \{\mathbf{x}, w, h\}$, a weight probability distribution is calculated in equation 10 from all samples indexed by i. The training samples are composed of:

- the positive (object) samples defined by $\lambda = 0$ and labelled $c(i) = 0$
- the negative (non-object) samples defined by $\lambda = 1$ and labelled $c(i) = 1$.

$$G^{(t)}_{\chi,\lambda}(b) = \sum_i w(i)\ \delta_\chi(i,b)\ (1 - |c(i) - \lambda|) \tag{10}$$

$$\delta_\chi(i,b) = \begin{cases} 1 & \text{if} \quad f_\chi(i) = b \\ 0 & \text{else} \end{cases} \tag{11}$$

In this equation 10 feature $f_\chi(i)$ represents in these works either the $SLBP_\chi$ or CCR_χ feature described above. The initial sample weights $w(i)$ are normalised over the distribution of samples ($\sum_i w(i) = 1$). A classifier $h^t_\chi(i)$ at iteration t is associated with each feature position χ in the samples template area. It classifies samples according to their feature response and their weight:

$$h^{(t)}_\chi(i) = \begin{cases} 1 & \text{if} \quad G^{(t)}_{\chi,0}(f_\chi(i)) > G^{(t)}_{\chi,1}(f_\chi(i)) \\ 0 & \text{else} \end{cases} \tag{12}$$

The Adaboost weak classifier $h^{(t)}_{\chi_m}(i)$ at iteration t is chosen at the feature position χ_m where the classifier is the most discriminative among all other classifiers $h^{(t)}_\chi(i)$ as follows:

$$\chi_m = \text{argmin}_\chi\{e_\chi\} \tag{13}$$

$$e_\chi = \sum_i w(i)\ |h^{(t)}_\chi(i) - c(i)| \tag{14}$$

In this equation the error term e_χ represents the sum of sample weights of the badly classified samples at feature location χ. Sample weights are then updated in the following equation where Z represents a normalising factor:

$$\alpha^{(t)} \ = \ \frac{1}{2} ln \frac{1 - e_{\chi_m}}{e_{\chi_m}} \tag{15}$$

$$w(i) \longleftarrow \frac{w(i)}{Z} \begin{cases} e^{-\alpha^{(t)}} & \text{if } h^{(t)}_{\chi_m}(i) = c(i) \\ e^{+\alpha^{(t)}} & \text{else} \end{cases} \tag{16}$$

Objects are detected as positive candidate using the trained data $G^{(t)}_{\chi_m,\lambda}(b)$ and $\alpha^{(t)}\ \forall t$ if equation 17 is verified.

$$H(i) = \sum_t \alpha^{(t)} h^{(t)}_{\chi_m}(i) > 0 \tag{17}$$

4 People, head and face detection

An object is detected when its SLBP features pass the Adaboost cascade of strong classifiers defined in equation 17. Objects of different sizes in an image are detected by a varying size scanning window across an image. This window has a user defined minimum size and a maximum size set by the image dimension. For instance, we reach 1fps for a minimum size of 160 pixels for people height in PAL images and when the scan is performed across images every N pixels horizontally and vertically where $N = 10\%$ of the scanning window width and height respectively. Integral images are calculated beforehand to speed up the process. Simple rules are applied to the detected candidates in order to fuse overlapping candidates and eliminate stand alone noisy candidates: objects are merged if they overlap each other with a minimum union-over-intersection ratio of 50% and a final object requires a minimum number of 2 overlapping ones.

An illustration of tracked faces, heads and people is given in figures 2. We have used a simple temporal window analyser [anonymous] to extract people trajectories.

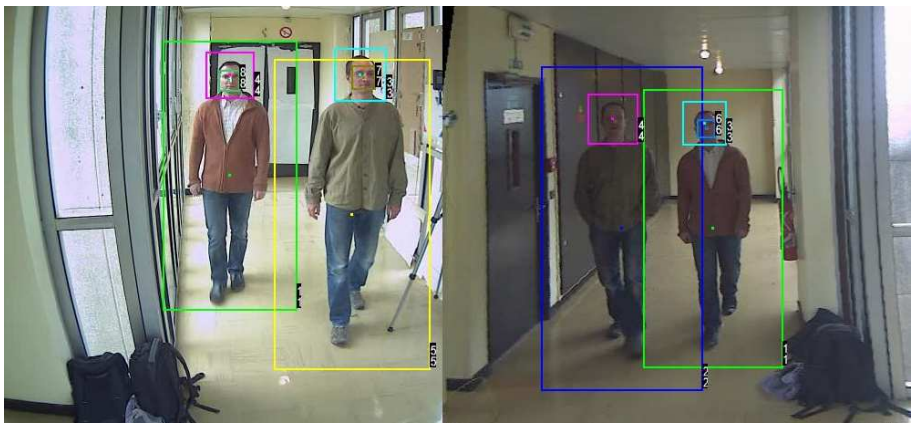

Fig. 2. Examples of tracked people, head and faces

5 Evaluation

5.1 People detection evaluation

We have evaluated our people detection algorithm on the test human dataset provided by INRIA against state of the art algorithms which we refer as HOG [4] and LBP-HOG [24]. The INRIA human dataset is composed of 1132 human images and 453 images of background scenes containing no humans. The results are displayed in figure 3 which shows that we obtain slightly better performances

than the HOG-LBP technique in terms of missed detection rate vs. FPPI i.e. False Positive Per Image. In this figure, two extreme functioning modes could be chosen: approximately 2 noisy detections are obtained every 1000 background images for 50% true positive detections or 1 noisy detection every 2 frames for a detection rate of approximately 88%.

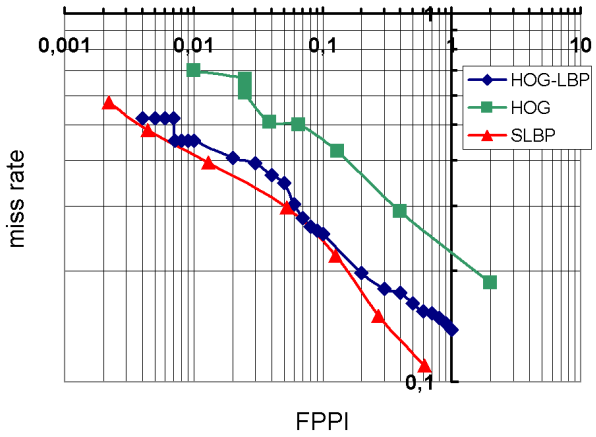

Fig. 3. People detection evaluation: False Positive Per Image vs. miss detection rate for the INRIA test database

5.2 Face detection evaluation

The same evaluation scheme of people detection above is used for face detection. The FPPI rates are obtained on 997 NICTA background images of 720x576 pixels. The scanning is performed for minimum face sizes of 20 pixels (and maximum the image height) at every 10% of its actual sizes as described earlier. 180 faces provided by a CMU test face image database are used to evaluate true positive rates. We have compared our results with the 2 versions of Haar feature provided by the OpenCv library i.e. the standard 'default' and alternative 'alt' training parameters. The results in table 1 show that the Haar 'alt' technique performs better than the traditional Haar one. And our CCR technique provides similar face detection rates while giving a less false alarm rate. The proposed approach is approximately 1% less successful in detecting faces than the Haar technique while this latter is 32% more noisier than our CCR technique.

6 Conclusion and future works

We have here presented a novel couple cell response feature for face detection. A simplified LBP feature is proposed to detect people and head. The Adaboost

technique	TP(%)	FPPI
Haar (default)	91.57	4.132
Haar (alt)	92.13	1.685
CCR	91.01	1.274

Table 1. Face detection evaluation

training scheme is adopted to train these features. The evaluation shows that we obtain state of the art performance for people and face detection. These features shall be studied for various body parts to allow algorithms to choose the best appropriate feature to detect people successfully. Unlike traditional feature kernel which provide binary output between 2 cells of pixels (e.g. Haar and LBP), the couple cell response study the amplitude of pixel distribution comparison. To give a more discriminative detection power, databases shall be revised for such features. For instance, more samples and more difference in lighting contrasts should influence the detection rates.

Acknowledgements

We would like to thank the ANR project 'VideoId' who partially founded this work and the vision group of the Telecom Sud Paris university for providing the new video dataset. However, the views and opinions expressed herein do not necessarily reflect those of the financing institutions.

References

1. Adam, A., Rivlin, E., Shimshoni, I.: Robust fragment-based tracking using integral histogram. In: Computer Vision and Pattern Recognition - CVPR (2006)
2. CAVIAR: Context Aware Vision using Image based Active Recognition. In: http://www.homepage.inf.ed.ac.uk/rbf/CAVIAR/. `http://www.homepage.inf.ed.ac.uk/rbf/CAVIAR/`
3. Chau, D.P., Bremond, F., Corvee, E., Thonnat, M.: Repairing people trajectories based on point clustering. In: In the International Conference on Computer Vision Theory and Applications (VISAPP) (2009)
4. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: Computer Vision and Pattern Recognition - CVPR (2005)
5. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: a benchmark. In: CVPR (2009)
6. Freund, Y., Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences (1997)
7. Huang, C., Nevatia, R.: High performance object detection by collaborative learning of joint ranking of granules features. CVPR10 - IEEE Conference on Computer Vision and Pattern Recognition (2010)
8. Hussein, M., Porikli, F., Davis, L.: Object detection via boosted deformable features. IEEE International Conference on Image Processing (ICIP) (2009)

9. Jordao, L., Perrone, M., Costeira, J., Santos-Victor, J.: Active Face and Feature Tracking. In: International Conference on Image Analysis and Processing - ICIAP (1999)
10. Laptev, I.: Improvements of object detection using boosted histograms. In: Proceedings of the British Machine Vision Conference (2006)
11. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 2, 91–110 (2004)
12. McKenna, S., Gong, S., Collins, J.: Face Tracking and Pose Representation. . In: BMVC (1996)
13. Mikolajczyk, K., Leibe, B., Schiele, B.: Multiple object class detection with a generative model. CVPR (2006)
14. Mikolajczyk, K., Schmid, C., Zisserman, A.: Human detection based on a probabilistic assembly of robust part detectors. In: ECCV (2004)
15. Mohan, A., Papageorgiou, C., Poggio, T.: Example-based object detection in images by components. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 349–361 (2001)
16. OpenCv: Intel Open Source Computer Vision Library. In: http://sourceforge.net/projects/opencvlibrary/. `http://sourceforge.net/projects/opencvlibrary/`
17. Schwerdt, K., Crowley, J.: Robust Face Tracking using Color. In: 4th IEEE International Conference on Automatic Face and Gesture Recognition (2000)
18. Singh, V.K., Wu, B., Nevatia, R.: Pedestrian tracking by associating trackets using detection residuals. IEEE Workshop on Motion and video Computing (WMVC) pp. 1–8 (2008)
19. Suguna, R., Sudha, N., Sekhar, C.: A fast and efficient face detection technique using support vector machine,. N.R. Pal et al. (Eds.): ICONIP 2004, LNCS 3316 (2004)
20. Trefny, J., Matas, J.: Extended Set of Local Binary Patterns for Rapid Object Detection. In: Computer Vision Winter Workshop 2010 - CVWW10 (February 2010), `http://cmp.felk.cvut.cz/cvww2010/`
21. Tuzel, O., Porikli, F., Meer, P.: Pedestrian detection via classification on riemannian manifolds. PAMI 30(10) (2008)
22. Varona, J., Buades, J., Perales, F.: Hands and face tracking for VR applications. In: Computers and Graphics. pp. 179–187 (2005)
23. Viola, P., Jones, M.: Robust real-time face detection. In: International Journal of Computer Vision (2004)
24. Wang, X., Han, T., Yan, S.: An hog-lbp human detector with partial occlusion handling. ICCV09 - International Conference on Computer Vision (2009)
25. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. CVPR pp. 951–958 (2006)
26. Yang, M.H., Kriegman, D.J. Narendra, A.: Detecting Faces in Images: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(1), 34–58 (2002)
27. Zhu, Q., Avidan, S., Yeh, M., Cheng, K.: Fast human detection using a cascade of histograms of oriented gradients. In: Computer Vision and Pattern Recognition - CVPR (2006)