Combining Rectangular and Triangular Image Regions to Perform Real-Time Face Detection

Hugo Proença and Sílvio Filipe
Dep. of Computer Science, Soft Computing and Image Analysis Group
IT - Networks and Multimedia Group
Universidade da Beira Interior, Covilhã, Portugal
Email: {hugomcp,sfilipe}@di.ubi.pt

Abstract

Nowadays, face detection techniques assume growing relevance in a wide range of applications (e.g., biometrics and automatic surveillance)and constitute a prerequisite of many image processing stages. Among a large number of published approaches, one of the most relevant is the method proposed by Viola and Jones [18] to perform real-time face detection through a cascade schema of weak classifiers that act together to compose a strong and robust classifier. This method was the basis of our work and motivated the key contributions given in this paper. At first, based on the computer graphics concept of "triangle mesh" we propose the notion of "triangular integral feature" to describe and model face properties. Also, we show results of our face detection experiments that point to an increase of the detection accuracy when the triangular features are mixed with the rectangular in the candidate feature set, which is considered an achievement. Also, it should be stressed that this optimization is obtained without any relevant increase in the computational requirements, either spacial or temporal, of the detection method.

Keywords*: Face Detection, Feature Extraction, Object Detection.*

1. Introduction

Object detection, and in particular, face detection is an important stage of various computer vision tasks, such as image retrieval, shot detection, video surveillance, etc. The goal is to locate objects of a pre-defined class in an image. For complex objects, such as human faces, it is hard to find features and heuristics able to perceive the varying appearance that object instances often have (e.g., faces may be rotated in all three directions, some people wear glasses, moustaches, beards, portions of the face are often occluded, etc). On these challenging conditions it is usual to apply statistical models previously trained to detect the desired type of objects. Statistical pattern recognition approaches rely on multiple instances of the desired object (*positive* samples) and of any other types of data (*negative* samples) to build a training set, from which the discriminating model is further inferred. Among multiple proposals, one of the most successful is the Viola and Jones' method [18] to perform real-time face detection.

In this paper, we propose a new type of features (*triangular integral features*) which, similarly to those originally proposed be the aforementioned authors, are computed in a single image scan. Our hope is that triangular regions better adapt to the face morphology and increase the final classifier discriminating capacity between *face* and *non-face* regions. The underlying concept behind this strategy is the notion of *triangle mesh*, widely used in the computer graphics domain to model surfaces (e.g., [16], [13] and [11]). A triangle mesh comprises a set of triangles (usually in three dimensions) that are connected by their common edges. Typically, computer graphics do operations on the vertices at the corners of triangles. On an individual triangle, an operation must be applied on the three vertices, while on large meshes there could be eight or more triangles meeting at a single vertex, being possible to operate on this vertex and achieve an identical effect with a fraction of the work.

1.1. Face Detection

The description of the face detection task is simple: given an image, the goal is to discriminate between regions that correspond to human faces and those that correspond to any other type of data. However, there are several factors (e.g., pose, occlusions and imaging conditions) that significantly increase the challenges of this apparently simple task.

There are - at least - three relevant face detection surveys. The former was published by Samal and Iyengar [15] in 1991. Few years later, Chellapa *et* al [3] published their work and, more recently, Yang *et* al. [20] deepen the level of analysis and published an extensive resume of the most relevant approaches. According to the authors of the later work, image detection methods can be divided into four categories, although some published approaches overlap more than a category:

1. Knowledge-based methods. These type of approaches are generally expert systems that encode the human sensibility about the patterns of a typical human face. Typically, several relevant regions are detected (eyes, nose, mouth) and the relationship between these features analyzed to decide wether they compose a face or not. Regarding this category, a former rule-based hierarchical method proposed by Yang and Huang [19] motivated later optimized and more successful approaches, as the work of Kotropoulos and Pitas [10].

2. Feature invariant approaches. According to the structural pattern recognition paradigm, these methods aim to find the underlying face structure that remains stable, even with varying poses, viewpoints or lighting conditions. Using a morphology-based technique to extract "eye-analogue" segments, Han *et* al. proposed a structural face detection method [8] that acted as basis for various posterior approaches.

3. Template matching methods. In order to characterize the facial appearance, these type of methods usually store several face patterns, either as a whole or by separate sub-regions. Later, the correlation between an input image and the stored templates gives the probability of a region to contain a human face. This category comprises the largest number of published approaches, as more closely fits the intuitive statistical pattern recognition paradigm.

4. Appearance-based methods. Oppositely to template matching techniques, appearance-based models use machine learning techniques to extract representative facial features, often difficult to perceive by humans. The learned models are further used for classification purposes, either through discriminant functions (e.g. [17]), or nonlinear decision surfaces (e.g., [14]). Regarding the key feature extraction stage, texture-based features are often used, through second-order statistical moments (e.g., Augusteijn and Skufca [1]). Starting from RGB images, Crowley and Berard [5] computed the Red and Green components histograms to obtain the probability of a particular RGB vector, given that the pixel observes skin. The utilization of Gaussian density functions and of mixture of Gaussians should also be mentioned, usually to model skin color and act as relevant features (e.g., [2]). Viola and Jones [18] used a large set of "weak features", each one consisting exclusively on the average difference between rectangular image regions, combined through an ensemble techniques into a strong and very fast classifier. As this approach is the basis of the method proposed in this paper, it is detailed in section 2.

The remainder of this paper is organized as follows: section 2 briefly describes the method proposed by Viola and Jones [18] to perform real-time face detection. A detailed description of the proposed detection strategy is given in section 3. Section 4 reports our experiments and discusses the results and, finally, section 5 gives the conclusions and points some directions for our further work.

2. Viola and Jones' Method

In [18] authors describe a very fast face detection framework, able to real-time processing with impressive low error rates. The key idea behind their method is the notion of *integral image*, that is built with a single image scan and supports the feature extraction process. Starting from an intensity image i with dimensions $W \times H$, the integral image ii has the same dimensions of i and, for a pixel (x, y), is simply given by the sum of the intensities of the pixels located above and to the left

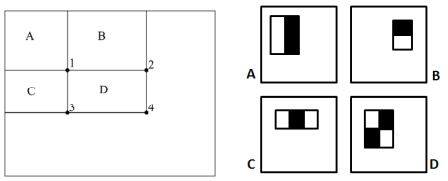
$$ii(x, y) = \sum_{c=1}^{x} \sum_{r=1}^{y} i(c, r) \tag{1}$$

Due to computation concerns, authors proposed a pair of recurrences that enable the calculus of the integral image through an unique image scan:

$$s(x, y) = s(x, y - 1) + i(x, y) \tag{2}$$
$$ii(x, y) = ii(x - 1, y) + s(x, y) \tag{3}$$

where $s(x, y)$ is a cumulative row sum and, by definition, $s(x, 0) = 0$ and $ii(0, y) = 0$.

There are some interesting properties of integral images. At first, as illustrated by figure 1a, the sum of any region can be computed using simple arithmetic of four array references. The area of region D is given by $ii(x_4, y_4) + ii(x_1, y_1) - (ii(x_2, y_2) + ii(x_3, y_3))$, where (x_j, y_j) denotes the position of the pixels with label j. Also, the invariance of this method to scale should be highlighted, as the computation of the area of large regions requires an equal number of operations as small ones. Obviously, this is one of its key properties, regarding the ability of real-time data processing.



(a) (b)

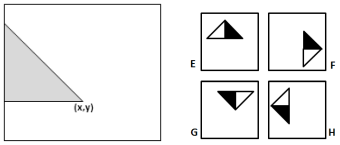
Figure 1: Computation of the average intensity of an image region through the notion of integral image (a) and type of rectangular features (A, B, C and D) originally proposed by Viola and Jones (b).

Hereinafter, the problem is to learn a classification function that, using a small feature subset, is able to distinguish between *face* and *non-face* regions. Authors used a large images training set, similar to those illustrated in figure 4. They considered each of the 160 000 candidate features as a potential weak classifier that simply computes the value of the respective feature and outputs a binary decision (*face* or *not-face*), wether the value is higher or lower than a threshold. Further, an AdaBoost algorithm is used to combine k weak classifiers into a final strong one. This learning algorithm iterates k times over the complete training set and selects one weak classifier per iteration. At iteration j, the selected weak classifier is the one that minimizes the error rates on the instances where the $j-1$ previously selected weak classifiers tend to fail.

As we confirmed in our experiments, this learning strategy constructs strong classifiers that operate remarkably fast and with good generalization capabilities. Nevertheless, authors proposed a final optimization step for better performance. Instead of using one single strong classifier, which typically analyzes about 80 features to achieve good accuracy, they proposed the use of an attentional cascade of small strong classifiers, each one analyzing less than five features. The key insight is that very simple classifiers (placed on the beginning of the cascade) are able to reject the large majority of negative sub-windows, while accepting all the positives ones and acting as filters that guarantee that slightly more complex classifiers located at the end of the cascade only analyze image windows with high probability of containing a face. When compared with one single strong classifier, authors concluded that this classification cascade has better performance at the expenses of a slight deterioration in accuracy.

3. Proposed Method

As above mentioned, the underlying idea of the proposed feature set cames from the notion of *triangle mesh*. Our hope is that bidimensional triangular features are able to model the human face morphology better than the the original set of rectangular features. At least, we hope that our types of features and those originally proposed by the authors can complement themselves to increase the discriminating capacity of the final strong classifier. In this section, we define and illustrate the four types of proposed features and give an algorithm to compute them through an unique image scan, which we considered essential to maintain the capacity to deal with real-time data.



(a) (b)

Figure 2: Notion of triangular integral value at pixel (x, y) (figure (a)) and type of proposed features (figure (b)).

3.1. Triangular Integral Features

Figure 2b illustrates the four types of proposed features to model human faces, from now on called E, F, G and H to distinguish them from the original types of features. Similarly to the proposal of Viola and Jones, every feature is given by the difference of the sum of intensities of the dark and light regions, whose we propose to be rectangular triangles. Thus, features E, F, G and H are equilateral triangles with four different orientations (respectively top, left, down and right), being each one divided into two rectangular triangles (dark and light sub-regions of the triangles of figure 2b). These can be of four types, whose we labeled respectively *North-West (nw)*, *North-East (ne)*, *South-West (sw)* and *South-East (se)* due to the image corner adjacent to the triangle rectangular angle. As an example, the E triangular feature is divided into a se (light sub-region) and a sw (dark sub-region) rectangular triangle.

For a base resolution of 24×24 windows and allowing varying sizes and locations for each type of features it is possible to build an exhaustive set of 76 000 features, all of them based on the notion of *triangular integral image* (ti). Figure 2a illustrates the sw triangular integral at point (x, y), which sums the intensities of the pixels contained within the dark triangular region. Formally, let i be an intensity image with H rows and W columns and $i(x, y)$ be the intensity value of the pixel located at column x and row y. The four types of triangular integral images are given by

$$ti_{sw}(x,y) = \sum_{r=1}^{x}\sum_{c=1}^{r} i(c, y-x+r) \quad (4)$$

$$ti_{se}(x,y) = \sum_{r=1}^{W-x}\sum_{c=1}^{r} i(W-c, y-W+x+r) \quad (5)$$

$$ti_{nw}(x,y) = \sum_{r=1}^{x}\sum_{c=1}^{r} i(c, y+x-r+1) \quad (6)$$

$$ti_{ne}(x,y) = \sum_{r=1}^{W-x}\sum_{c=1}^{r} i(W-c, y+W-x-r) \quad (7)$$

considering, by definition, $i(x,y) = 0$ if $x <= 0$ or $y <= 0$ or $x > W$ or $y > H$.

3.2. Calculus of the Triangular Integral Images With a Single Image Scan

In this section we describe four pair of recurrences that allow the computation of the integral images, either the original rectangular and our proposals, in a single image scan. As it follows from (2)-(7), the computation of each feature requires previously known neighbor values. Thus, to compute the original integral image and our proposed sw triangular integral, an image must be scanned from top to down and from left to right order. The remaining types of integral images simply require the image scan in different order.

ne Triangular Integral Images

$$s_{ne}(x,y) = s_{ne}(x,y-1) + I(x,y) \quad (8)$$
$$ti_{ne}(x,y) = ti_{ne}(x-1,y-1) + s_{ne}(x,y) \quad (9)$$

nw Triangular Integral Images

$$s_{nw}(x,y) = s_{nw}(x,y-1) + i(x,y) \quad (10)$$
$$ti_{nw}(x,y) = ti_{nw}(x+1,y-1) + s_{nw}(x,y) \quad (11)$$

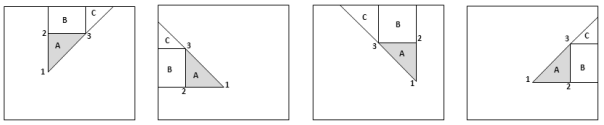
se Triangular Integral Images

$$s_{se}(x,y) = s_{se}(x,y+1) + i(x,y) \quad (12)$$
$$ti_{se}(x,y) = ti_{se}(x-1,y+1) + s_{se}(x,y) \quad (13)$$

sw Triangular Integral Images

$$s_{sw}(x,y) = s_{sw}(x,y+1) + i(x,y) \quad (14)$$
$$ti_{sw}(x,y) = ti_{sw}(x+1,y+1) + s_{sw}(x,y) \quad (15)$$

Based on the afore described types of integral images, it is possible to compute the triangular features using simple array references, as illustrated in figure 3. For instance, the triangular area labeled with "A" in figure 3a is given by $ti_{nw}(1) - (ii(3) - ii(2) - ti_{nw}(3))$. The remaining types of triangular features are similarly obtained. According to this process, it is possible to obtain the values of all the proposed features, as they combine pairs of the described four types of rectangular triangles (figure 2).



(a) (b) (c) (d)

Figure 3: Computation of the triangular integral features proposed in this paper, based in the notions of integral image originally published by Viola and Jones [18] and in our proposed triangular features. The computation of the average intensity of the regions labeled with "A" on the upper figures is respectively based on the nw (figure 3a), sw (figure 3b), ne (figure 3c) and se (figure 3d) triangular integrals images.

4. Experiments and Discussion

To evaluate the merits of the proposed feature set some experiments were carried out. This section describes them, namely the used data sets and evaluation measures. Here, we regard the problem of face detection as a binary classification task, where for each image region the system outputs one of two possible decisions: (*face* or *non-face*). Consequently, two common types of errors (false positives and false negatives) can be obtained. These types of errors allow the obtainance of the false positive and false negative rates, as well of the approximated equal error rates.

Figure 4: Example of the face images used in our experiments. These were extracted from the web and contain faces from caucasian, asian and black people with varying poses, lighting conditions and backgrounds.

Similarly to the evaluation process described by Viola and Jones, both the learning of the classifier functions and their consequent evaluation were performed in a large set of images (30 000 images in our case, divided into 10 000 faces and 20 000 non-faces) randomly extracted from the world wide web, cropped and resized to dimensions of 24×24 pixels. The face images have varying poses and were acquired under very heterogeneous lighting conditions and imaging frameworks. Also, the data set contains im-

Type of Features	Proportion (%) Regarding the Selected Feature Set
E	7.43
F	2.31
G	6.91
H	2.27
E, F, G and H	18.93%

Table 1: Average proportion of features of types E, F, G and H automatically selected by the boosting learning algorithm to be included in the final strong classifier.

ages of people with different skin pigmentation, either caucasian, black, oriental or indian subjects. Backgrounds vary from very simple to high complex, in order to increase the difficulty of the detection task. Non-faces images are from heterogeneous scenarios, either buildings, nature, roads, sky and many other types of textures. Figure 4 illustrates some of the face images used in our experiments. It should be stressed that although we used faces with fixed dimensions for evaluation purposes, the given method could easily be applied to detect faces with different sizes, as described in the work of Viola and Jones [18].

All the results that are given below are averages of twenty learning and classification processes, using an hold-out validation scheme. At each iteration, we randomly selected one third of the images to perform evaluation and used all the remaining images in the respective learning process.

4.1. Learning Classification Functions

In our first experiment we extracted the complete candidates feature set (of the original types A, B, C and D together with our proposals E, F, G and H) and analyzed the number of triangular features that are automatically selected by the boosting learning process. As before described, the Adaboost algorithm iteratively selects the feature which correspondent weak classifier minimizes the error rates on the instances that were incorrectly classified by previously selected classifiers. Thus, any feature of type E, F, G or H selected in this process can be regarded as an optimization to the classification process, as it was considered that, at a given learning iteration, was the one that better ensemble with weak classifiers previously selected.

Table 1 gives the average proportion of triangular features that were selected in our boosting learning experiments. It can be observed that, averagely, almost 20% of the candidate features selected for the inclusion in the final strong classifier were from the types proposed in this paper, which is a significant proportion. Also, we observed that features of types E and G - that correspond to the top and down oriented pair of triangles - were more frequently included in the final strong classifier, which we justify by their ability to respectively model the regions correspondent to the human head and mouth.

In our next experiment we analyzed the detection error rates obtained with and without our feature set proposal, which will give an idea of the amount of discriminating information added to the final strong classifier.

4.2. Classification Errors

We compared the error rates obtained by strong classifiers exclusively learned from the original feature set (features of types A, B, C and D) and together with our proposed candidates feature set (E, F, G and H). Figure 5 shows the obtained Receiver Operating Curves. The vertical axis give the sensibility (true positives rate) and the horizontal axis the complement of specificity (false positives rate). The original types of features are represented through the dotted lines and the union of these and our proposal by the continuous lines, being the results obtained using ten-fold cross validation. The benefits of adding the proposed types of features (E, F, G and H) to the candidates feature set of the strong classifier are evident, as it averagely decreased the recognition error rates by 4.32%. This decrease is specially relevant due to the very low error rates that the original classifier of Viola and Jones already achieve. Also, the obtained equal error rates (1.517% on the original proposal and 1.420% on ours) confirm the utility of the proposed candidates feature set.

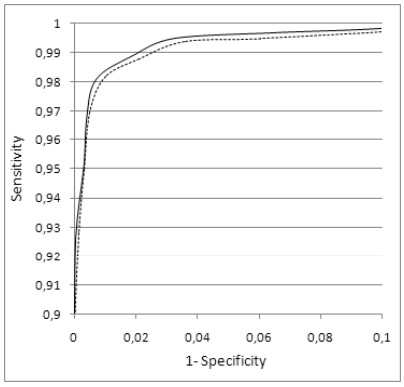


Figure 5: Comparison between the error rates obtained by strong classification functions learned from the original feature set proposed by Viola and Jones (dashed line) and from the union of these features and our proposals (continuous line).

5. Conclusions

Among multiple proposals, one of the most known and practically face detection methods used for commercial purposes is the approach of Viola and Jones [18], which uses the notion of integral image to compute a set of rectangular features that proved to effectively distinguish between *face* and *non-face* image regions. Based on the concept of *triangular mesh*, widely used in the computer graphics domain for the purpose of surface modeling, we proposed a set of bidimensional triangular features and gave an algorithm to compute them through a single image scan, due to real-time processing concerns. In the experiments we used the complete feature set (Viola and Jones' original and our proposals) and observed that the boosting learning method automatically selected a significant portion of the proposed features for the final strong classifier, which is a good indicator of their utility. Later, we used a set of 30 000 images randomly extracted from the web to learn different classification functions, when starting exclusively from the feature set originally proposed and from the original together with our proposal. Results showed a decrease in the obtained Equal Error Rates of about 4.32% when both feature sets are used as candidate features, which we considered an achievement, specially due to the low error rates that the original work already achieved. Finally, it should be highlighted that this improvement in accuracy was obtained without any significant increase in the computational requirements. Forthcoming work is focused on the extraction of a set of triangular features with arbitrary rotation, again with a single image scan. We hope that this new set of features, as it more closely fits the notion of bidimensional triangular mesh, provides even better accuracy, while maintaining the computation requirements.

Acknowledgements

We acknowledge the financial support given by the Portuguese "FCT-Fundação para a Ciência e Tecnologia" and the European "FEDER" in the scope of the PTDC/EIA/69106/2006 research project "BIOREC: Non-Cooperative Biometric Recognition".

References

[1] M. Augusteijn and T. Skujca. Identification of human faces through texture-based feature recognition and neural network technology. In *Proceedings of the IEEE International Conference on Conf. Neural Networks*, pages 392–398, January 1993.

[2] J. Cai, A. Goshtasby, and C. Yu. Detecting human faces in color images. In *Proceedings of the 1998 International Workshop Multi-Media Database Management Systems*, pages 124–131, 1998.

[3] R. Chellappa, C. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proceedings of the IEEE, vol. 83, no. 5*, pages 705–740, 1995.

[4] I. Craw, H. Ellis, and J. Lishman. Automatic extraction of face features. *Pattern Recognition Letters, vol. 5*, pages 183–187, 1987.

[5] J. Crowley and F. Berard. Multi-modal tracking of faces for video communications. In *Proceedings of the IEEEConference on Computer Vision and Pattern Recognition*, pages 640–645, 1997.

[6] Y. Dai and Y. Nakano. Face-texture model based on sgld and its application in face detection in a color scene. *Pattern Recognition, vol. 29, no. 6*, pages 1007–1017, 1996.

[7] V. Govindaraju. Locating human faces in photographs. *International Journal of Computer Vision, vol. 19, no. 2*, pages 129–146, 1996.

[8] C. C. Han, H. Liao, K. Yu, and L. Chen. Fast face detection via morphology-based pre-processing. In *Proceedings of the Ninth International Conference on Image Analysis and Processing*, pages 469–476, 1998.

[9] T. Jebara and A. Pentland. Parameterized structure from motion for 3d adaptive feedback tracking of faces. In *Proceedings of the IEEE International Conference Computer Vision and Pattern Recognition*, pages 144–150, The Netherlands, 1997.

[10] C. Kotropoulos and I. Pitas. Rule-based face detection in frontal views. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, vol. 4*, pages 2537–2540, 1997.

[11] T. Moller and E. Haines. *Real-Time Rendering*. A.K. Peters, U.S.A., 2002.

[12] C. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Proceedings of the International Conference on Computer Vision*, pages 555–562, January 1998.

[13] J. Rossignac. Edgebraker: Connectivity compression for triangle meshes. *IEEE Transactions on Visulization and Computer Graphics, Vol. 5, no. 1*, pages 47–61, 1999.

[14] H. Rowley, S. Baluja, , and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, no. 1*, pages 23–38, 1998.

[15] A. Samal and P. Iyengar. Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern Recognition, vol. 25, no. 1*, pages 65–77, 1992.

[16] Y. Sun, D. Page, J. Paik, A. Koshan, and M. Abidi. Triangle mesh based edge detection and its application to surface segmentation and adaptive surface smoothing. In *International Conference on Image Processing (ICIP'02)*, pages 825–828, June 2002.

[17] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience, vol. 3, no. 1*, pages 71–86, 1991.

[18] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision, vol. 57, no.2*, pages 137–154, 2004.

[19] M.-H. Yang, D. J. Kriegman, and N. Ahuja. Human face detection in complex background. *Pattern Recognition, vol. 27, no. 1*, pages 53–63, January 1994.

[20] M.-H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *Transactions on Pattern-Analysis and Machine Intelligence, vol. 24, no. 1*, pages 34–58, January 2002.