AN INTEGRATED APPROACH TO CONTEXTUAL FACE DETECTION

Santi Seguí[1], Michal Drozdzal[1,2], Petia Radeva[1,2] and Jordi Vitrià[1,2]
[1]*Computer Vision Center, Universitat Autònoma de Barcelona, Bellaterra, Spain*
[2]*Dept. Matemàtica Aplicada i Anàlisi, Universitat de Barcelona, Barcelona, Spain*

Keywords: Face detection, Object detection.

Abstract: Face detection is, in general, based on content-based detectors. Nevertheless, the face is a non-rigid object with well defined relations with respect to the human body parts. In this paper, we propose to take benefit of the context information in order to improve content-based face detections. We propose a novel framework for integrating multiple content- and context-based detectors in a discriminative way. Moreover, we develop an integrated scoring procedure that measures the 'faceness' of each hypothesis and is used to discriminate the detection results. Our approach detects a higher rate of faces while minimizing the number of false detections, giving an average increase of more than 10% in average precision when comparing it to state-of-the art face detectors.

1 INTRODUCTION

Although face detection is a classical computer vision problem addressed from the mid nineties(Rowley et al., 1995), its application to real world problems began with the publication of the Viola & Jones algorithm (Viola and Jones, 2001) in 2001. Since then, mass-market products such as Google's Picasa, Adobe's Photoshop Elements or Apple's iPhoto have made intensive use of this technology.

In spite of these successful applications, robust face detection in non controlled environments is still an open problem, as it has been recognized in several recent reports (Frome et al., 2009; Zhang and Zhang, 2010). The difficulty in developing a robust detector arises from the diversity of human faces (e.g. the variability in location, pose, orientation, and expression) and changes in environmental or acquisition conditions (like: resolution, illumination, exposure, occlusions) (Yang et al., 2002) as well as due to objects worn by persons that may occlude their face (e.g. glasses, scarf, etc.).

In principle, one the most promising approaches for increasing face detector performance is the use of contextual information, but this strategy has been seldom considered in the literature (Kruppa and Schiele, 2003; Mikolajczyk et al., 2004). The main novelty of this paper is the definition of a general framework for integrating, in a discriminative way, contextual information in the detection of faces. This framework allows a two-fold novelty contribution that clearly improves results of non-contextual detectors: (i) The definition of an integrated score, which optimally combines all content- and context-based image evidences about the "*faceness*" of an image window; (ii) The integration of complementary detectors, which generate supplementary hypothesis that clearly improves the object detection results. Furthermore, the architecture of our framework is extensible: new content-based or context-based detectors can be easily plugged in.

We experimentally show, using different databases of several thousands of images, that the use of context significantly improves both precision and recall of the detection, giving an average improvement of more than 10% in Average Precision (AP) when comparing the framework to classical content-based detectors (even more in the most challenging databases, such as the Pascal VOC2010 (Everingham et al.,) database, where the improvement in AP is of 16%).

The paper is organized as follows: In section 2, we consider our method in the context of previous works. In section 3, we explain the methodology of integrating content- and context-based face detectors as well as the score definition. In section 4, we present some experimental results. Finally, the paper finishes by conclusions in section 5.

Seguí S., Drozdzal M., Radeva P. and Vitrià J. (2012).
AN INTEGRATED APPROACH TO CONTEXTUAL FACE DETECTION.
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pages 90-97
DOI: 10.5220/0003746200900097

2 RELATED WORK

As previously commented, the breakthrough in face detection was caused by the work of Viola and Jones in 2001 (Viola and Jones, 2001), who used Haar-like features combined with a boosting-based learning scheme. Since 2001, there have been many works aimed to enhance the Viola and Jones' detector. For instance, different feature types have been considered, such as generic linear features, composite features, variations from Haar features, etc. In addition to exploring better features, improvements to the boosting algorithm have also been proposed, such as using different boosting schemes, reuse of intermediate results, etc [1].

Today, most of the state-of-the art reported systems, such as DCascade (Xiao et al., 2007) or SCascade (Bourdev and Brandt, 2005) are still based on the basic paradigm of Viola and Jones.

The Viola and Jones' algorithm is a clear example of a content-based face detector. Such detectors locate faces by classifying the content of a detection window, iterating over all positions and scales of the input image. As an alternative to content-based detectors, context-based object detectors rely on the information about the environment-object relationship to infer the face position and scale in the image. The advantage of this approach is that it is more permissive about the object appearance, since it is detected by exploiting environment-object relationships.

Lately, some interesting algorithms for human detection have been proposed. In (Felzenszwalb et al., 2010), a latent SVM is used to define a Discriminatively trained Part-based Model (DPM) that can be applied to person detection. Another interesting approach for person detection is the Poselet-Based Detector (PBD) presented in (Bourdev and Malik, 2009) where information about specific body parts is used in a probabilistic Hough framework to detect human bodies in unconstrained images. All these methods can be appropriately adapted to perform as context-based face detectors.

Recently, some authors have used the visual information surrounding an hypothesis in order to improve face detectors, usually as a post-processing procedure. In (Atanasoaei et al., 2010), Atanasoaei et al. present a hierarchical model which is built using the detection distribution around a target hypothesis to discriminate between false alarms and true detections. A different approach was proposed in (H. Takatsuka and Okutomi, 2007) to study the score distribution in both location and scale space. All these methods use the word *context* as a proxy for the output of a single face classifier around the hypothesis. In (Kruppa and Schiele, 2003), a method is presented that focuses on the role of local context as a predictive cue for computational face detection. Local context is implemented by an object detector which is trained with instances that contain the entire head, neck and part of the upper body of a person. In experiments on two large data sets, they find that using local context can significantly increase the number of correct detections, particularly in low resolution cases. However, the question of optimal integration of different detectors is not considered and no results are presented with respect to such an integration of the local context detector as an auxiliary cue for a content-based detector.

The application of both, content- and context-based face detectors suffer from the same problem: they are dealing with a non-symmetric problem, and to get a low false detection rate (which is a requirement), they lost a non-negligible part of true detections. The idea behind our system is that by integrating several content- and context-based face detectors tuned to work with a higher true detection rate than when used independently (see Fig. 1), we can keep the false detection rate at normal levels.

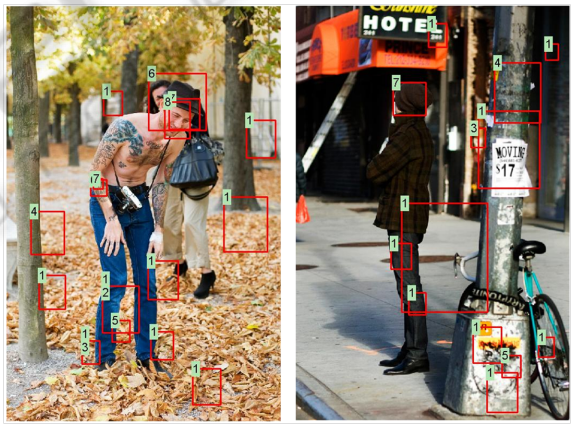


Figure 1: Content- and context-based detectors working at high true detection rates produce a lot of false positives which can be managed by integrating all detectors.

3 FACE DETECTOR SYSTEM

The proposed face detection system is divided in two main blocks. In the first stage, all possible face hypotheses are generated by the individual face detectors and a score which represents the detection confidence is assigned to each one.

In the second stage, the information from all hypotheses (hypotheses locations, scales and scores) is used in order to assign an integrated confidence score

[1] See (Zhang and Zhang, 2010) for an extensive report on all these variations.

to each detection. The optimal score assignment is learned from training data in order to maximize the integrated detector performance. To be able to compare hypotheses that are generated by different algorithms, the scores of each hypothesis should be properly normalized.

3.1 Hypotheses Generation and Scores

The goal of this stage is to generate all possible hypotheses of faces in the given image and to assign an initial score to each one. Our current implementation uses two different types of detectors: 1) five content-based detectors and 2) two contextual detectors.

3.1.1 Content-based Face Detectors

We have used the Viola and Jones face detector implemented in OpenCV 2.1[2], but any other content-based detector can be used. In our current implementation, we use five different cascades for face detection: four cascades trained for frontal face detection using different models and one cascade for detecting profile faces.

According to (Atanasoaei et al., 2010), the number of detections in the neighborhood of an hypothesis can be used as a reliable confidence score for this detector. In general, if the face is well defined in the image, it is highly probable that in several window locations, close to the real object location, we will get face detections. We have used this criterion to built a score for each face hypothesys generated by these detectors.

3.1.2 Context-based Face Detectors

Context-based face detectors assume that there is a high correlation between head position and body parts. So the knowledge about the body position and scale in the image can approximately determine face position and scale. Two state-of-the-art person detectors were modified in order to detect faces: 1) the Discriminatively Trained Part Based Model (DPM)[3] (Felzenszwalb et al., 2010) and 2) the Poselets Based Detector (PBD)[4] (Bourdev and Malik, 2009).

DPM method proposed by P. Felzenszwalb et.al. in (Felzenszwalb et al., 2010) is built on the pictorial structures framework. Pictorial structures represent objects by a collection of parts arranged in a deformable configuration. Each part captures local appearance properties of an object while the deformable configuration is characterized by spring-like connections between certain pairs of parts. DPM model was trained and as a result an upper body model was created. The obtained model has a component part which determines the head position. The hypothesis score is defined as the DPM response on the detected object.

The PBD method proposed by L. Bourdev et.al. in (Bourdev and Malik, 2009), is based on a large number of part detectors called "poselets". In the original method, every poselet votes for the location of body using the Max Margin Hough Transform (M2HT) (Maji and Malik, 2009). This approach was modified in order to obtain a vote for the head position. The voting scheme was learnt using the H3D database (Bourdev and Malik, 2009). The number of poselets was reduced from $\sim$ 1100 (original PBD method) to 250 by selecting the most informative ones about head position. Similarly as in (Bourdev and Malik, 2009), the hypothesis score is defined as the sum of the scores of the poselets which are voting for the same head position. We assume that the higher the sum of the score of all poselets voting the same detection, the higher the confidence about the face hypothesis.

The application of content- and context-based detectors in an arbitrary image normally results in a number of hypotheses distributed all over the image (see Fig.1). Let d_i be the set of available detectors. Let h^i_j be the face hypothesis j generated by d_i, which is defined by its location and scale in the image $(x_{i,j}, y_{i,j}, s_{i,j})$. Let $s_i(h^i_j)$ be the score produced by the detector d_i. Our objective is to define a score $S(x_{i,j}, y_{i,j}, s_{i,j})$ for each hypothesis location in the image which integrates the information from all detectors in an optimal way. The score function is defined on hypotheses locations and not on hypotheses because a location can be shared by several hypotheses from different detectors.

3.2 Merging Hypotheses

By combining hypotheses from different and complementary methods, two benefits are expected:

1) *Precision* improvement: the system should be able to generate detections with high confidence by analyzing the position of multiple detectors hypotheses; and

2) *Recall* improvement: faces that are not detected by one method can be detected by other approaches.

The merging procedure works in two steps. Firstly, there is a geometric merging of hypotheses

[2]Source code available at: sourceforge.net/projects/opencvlibrary/

[3]Source code available at: people.cs.uchicago.edu/$\sim$pff/latent/

[4]Source code available at: www.eecs.berkeley.edu/$\sim$lbourdev/poselets

produced by the same detector which have similar location and scales. Secondly, hypotheses from different detectors are merged.

3.2.1 Geometrical Merging

Given a set of hypotheses $\{h_j^i\}$ from detector d_i, we apply a weighted agglomerative clustering with pairwise distances based on location and scale similarity. A new hypothesis, defined by the mean location and scale from the set of merged hypotheses is created for every cluster. The process is independently repeated for all detectors. Figure 2(a) shows all face hypotheses produced by all detectors, and Figure 2(b) shows the hypotheses generated by the geometrical merging procedure.

3.2.2 Score Normalization

As commented before, an image location can be shared by multiple hypotheses from different detectors, which produce scores that are not directly comparable. This leads to the necessity for a score normalization algorithm.

In order to be able to merge hypotheses, a two-fold procedure is applied: 1) an hypotheses score normalization and 2) merging the hypotheses by considering detector scores. Both methods are based on the empirical performance of detectors in a training database.

Let $\mathbf{h}_j^i$ be a boolean variable corresponding to the hypothesis h_j^i which indicates if it is a true positive or a false positive. We use a score normalization procedure based on estimating the posterior probability of $\mathbf{h}_j^i$ given the detection score $s_i(h_j^i)$ of detector d_i:

$$p(\mathbf{h}_j^i|s_i(h_j^i),d_i) = \frac{p(\mathbf{h}_j^i|d_i)\cdot p(s_i(h_j^i)|\mathbf{h}_j^i,d_i)}{p(s_i(h_j^i)|\mathbf{h}_j^i,d_i)p(\mathbf{h}_j^i|d_i)+p(s_i(h_j^i)|\neg\mathbf{h}_j^i,d_i)p(\neg\mathbf{h}_j^i|d_i)} = \frac{p(s_i(h_j^i)|\mathbf{h}_j^i,d_i)}{p(s_i(h_j^i)|\mathbf{h}_j^i,d_i)+p(s_i(h_j^i)|\neg\mathbf{h}_j^i,d_i)} \quad (1)$$

where we assumed that $p(\mathbf{h}_j^i|d_i) = p(\neg\mathbf{h}_j^i|d_i)$, since we have a balanced dataset.

All terms in Eq.(1) can be easily and robustly estimated by using a large training image database. Then, if we assume that $p(s_i(h_j^i)|\neg\mathbf{h}_j^i,d_i)$ and $p(s_i(h_j^i)|\mathbf{h}_j^i,d_i)$ are exponentially distributed when $s_i(h_j^i)$ are on the wrong side of the detector threshold, we can apply the logistic regression assumption and compute a probability estimate of $p(\mathbf{h}_j^i|s_i(h_j^i),d_i)$ by applying, for example, the modified Platt's method (Lin et al., 2007).

3.2.3 Hypothesis Score Determination

Once all individual detector scores are properly normalized, we can proceed to the determination of the integrated score.

Let's define the *activation value* of an hypothesis location with respect to a detector d_i as:

$$a_i(x_{i,j},y_{i,j},s_{i,j}) = p(h_j^i|\mathbf{h}_j^i,d_i)p(\mathbf{h}_j^i|s_i(h_j^i),d_i) \quad (2)$$

The first term in Eq.(2) represents the *a priori* knowledge, we have about a specific detector and the second one, the probabilistic score of the detection.

Then, we can built the *activation vector* of an hypothesis location as $A(x_{i,j},y_{i,j},s_{i,j}) = [a_1(x_{i,j},y_{i,j},s_{i,j}),...,a_k(x_{i,j},y_{i,j},s_{i,j})]$, where k is the total number of detectors. This vector encodes all evidences from all detectors for a given image window.

In order to learn an optimal integrated score with respect to the distribution of $A(x_{i,j},y_{i,j},s_{i,j})$, we propose the definition of the score of an hypothesis location $S(x_{i,j},y_{i,j},s_{i,j})$ as a distance between its activation value vector $A(x_{i,j},y_{i,j},s_{i,j})$ and the decision surface that best discriminates the activation vectors corresponding to true positives from the activation vectors corresponding to false positives in the training database.

More specifically, we have defined the score of an hypothesis as the distance of its activation vector to the hyperplane defined by a RBF-Perceptron algorithm (Orabona et al., 2009) on the training data. The integrated score, for a new hypothesis location, takes the form:

$$S(x_{i,j},y_{i,j},s_{i,j}) = \sum_{k=1}^{N} \alpha_k K(A_k^{tr},A(x_{i,j},y_{i,j},s_{i,j})) \quad (3)$$

where $A_1^{tr},...,A_N^{tr}$ are the set of training activation vectors, $\alpha_1,...,\alpha_N$ are real weights estimated by the learning algorithm when minimizing the cumulative hinge-loss suffered over a sequence of examples and $K()$ is a kernel function.

4 EXPERIMENTS

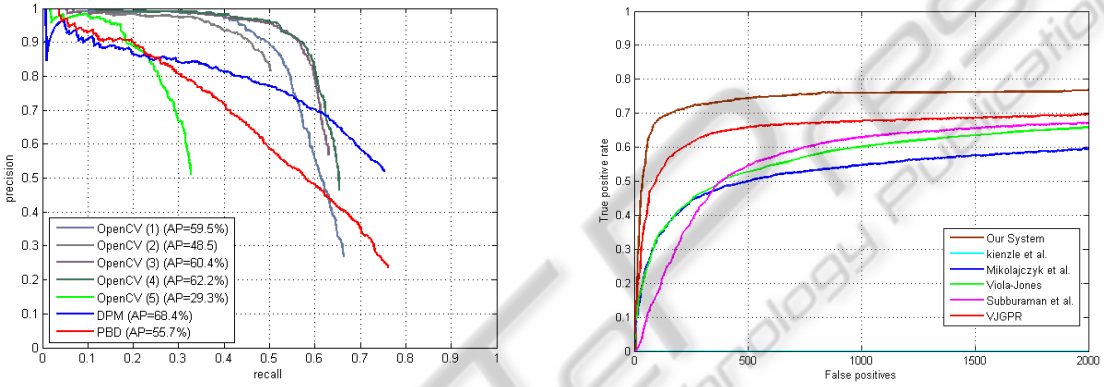
Databases. In order to test our algorithm, six different databases of faces in unconstrained conditions were used: 1) The 1000Portraits_BCN[5], 2) The Sartorialist[6], 3) The Pascal VOC 2010 Person Layout

[5]http://www.flickr.com/photos/1000portraitsbcn/sets/72157620858440104/

[6]http://thesartorialist.blogspot.com/

(a) (b)

Figure 2: A geometrical merging method, based on a weighted agglomerative clustering, can filter most of the redundant hypotheses from image (a) to produce a reduced set showed in image (b).



(a) (b)

Figure 3: a) Recall/precision curves for each individual content-based face detector when analyzing the Sartorialist database. (b) Results for the FDDB including our method, the Viola and Jones face detector (Viola and Jones, 2001), the method of Mikolajczyk et al. (Mikolajczyk et al., 2004), the method of Subburaman (Subburaman and Marcel, 2010), the method of Kienzle (Kienzle et al., 2004) and the method of Jain et al. (Jain and Learned-Miller, 2011).

database [7], 4) the CMU Face database [8], 5) The CMU Profile Face Images Database [9], and 6) The FDDB database [10].

The 1000Portraits_BCN database consists of 1000 images collected in streets of Barcelona; in most of the images only one person appears. The Sartorialist database also contains 1000 images taken in the streets of different cities and is characterized by the variety in both body pose and number of persons. The most challenging database is a set of images from validation set of Pascal VOC 2010 person layout challenge (Everingham et al.,), which contains 376 images depicting one or several people. The CMU Face database is classical set, which includes 130 images and 507 faces. The CMU Profile Face database contains 208 gray scale images with faces in profile views. Finally, the FDDB database (Jain and Learned-Miller, 2010) is the biggest face detection benchmark at this moment: 5171 faces in a set of 2845 images.

Evaluation Methodology. All visible faces in the databases were manually annotated. For the evaluation of the algorithm, the precision/recall curve is constructed using the confidence to determine the results ranking, accordingly to the Pascal VOC 2010 (Everingham et al.,) criterion (which is a *de facto* standard comparison measurement for object detection). For every precision/recall curve, the area under it, called Average Precision, is calculated. Face detections are considered True or False Positives based on the area of overlapping with ground truth bounding boxes. As correct detections, we consider the ones with the area of overlapping between the predicted bounding box

[7]http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2010/

[8]http://vasc.ri.cmu.edu//idb/html/face/frontal_images/index.html

[9]http://vasc.ri.cmu.edu//idb/html/face/profile_images/index.html

[10]http://vis-www.cs.umass.edu/fddb/

and ground truth bounding box $> 50\%$.

Training of the final version of our system has been performed using exclusively the Sartorialist database. Results for this database have been estimated by using a 10-fold cross validation. All other results for other databases are based on this training.

4.1 Experiments

First of all, in order to evaluate the benefits of our approach, we have evaluated the average precision of all individual face detectors used in the integrated approach for the Sartorialist database. The recall/precision curves of individual detectors are presented in figure 3(a). We can observe that the OpenCV detectors have very high precision for high confidence values. In contrast, the contextual detectors are characterized by a higher recall: they are able to detect more faces, but nevertheless their overall performance is quite poor. It is worth mentioning that the OpenCV(5), which is the profile face detector, obtains the worst AP value. Its behavior in the other databases is similar.

We also have evaluated the performance on several databases of a reduced version of our system: the one that uses exclusively content-based detectors. These results are useful to measure the relative importance of content- and context-based information in the full approach. These results are in the first column in Table 1. In all cases the contribution of context is quite important and produces an increment in AP that goes from 7% to more than 30%.

Table 1: Average Precision of the two versions of our system for different databases.

Database	*Content-based*	*Integrated*
1000Portraits_BCN	88.1%	95.4%
Sartorialist	68.0%	82.3%
Pascal 2010	50.8%	66.2%
FDDB	76.8%	79.4%
Sartorialist*0,25	33,5 %	65 %

In order to see the behavior of the method with low resolution databases, we performed an experiment with images from the Sartorialist database resized by the factor of 0.25. This database is composed of high resolution images with medium size faces. The results are presented in the fourth row of Table 1. In images with low face resolution, the content-based face detectors perform quite poor and the improvement when adding contextual information is remarkable (more than 30% improvement in AP).

We have also evaluated our system with the recently proposed FDDB database. The results in Figure 3(b) show that our system performs the best, with a large margin, with respect all reported results (our system gets an AP=79.4%, being the second best result the OpenCV implementation of the Viola & Jones algorithm with an AP=64.3%).

Finally, we have performed a set of experiments to compare our system to several state-of-the art methods that have reported results for the CMU-MIT Face database and the CMU Profile Face database: the DCascade and the SCascade methods (Xiao et al., 2007; Bourdev and Brandt, 2005). This comparison, shown in Table 2, is based on the reported results from these databases: the number of true detections when detecting the first false positive. For completeness, we also calculated our average precision for the CMU-MIT database, which is 94%.

Table 2: Performance comparison (true detection/false positive) for the CMU databases.

Method	*CMU-MIT Face*	*CMU Profile*
DCascade	440/1	108/1
SCascade	323/0	85/0
Our System	446/0	132/1

As can be seen in these tables and graphs, our method remarkably outperforms state-of-the-art methods on public databased. The contribution of context to these results can be fully appreciated when comparing the first column to the second column of Table 1: in all cases context information plays an important role.

With respect to the FDDB database, a challenging database with a large number of labeled faces, our method ranks the first with a clear advantage with respect to all reported methods. Nevertheless, results show that face detection is not yet a solved problem.

In order to give qualitative information about the behavior of our method, we have built a mosaic in Figure 4, where a sorted set of detected faces is shown. We can see that the scoring system assigns a low value to all false positives (red frames in the color figure). In most of the cases, these frames contain a face or head with a very bad estimation of scale or a head seen from behind. We can also see that most of the detected faces that would be missed by content-based detectors (green frames in the color figure) correspond to faces in unusual poses. Finally, we can also observe that the lowest scores are assigned to very low resolution faces.

5 CONCLUSIONS

We have presented a simple yet powerful approach

Figure 4: This figure shows a ordered set of detections of our system. The upper-left image is the highest score detection and the lower-right the one with the lowest score. Some detections have a color frame which indicates the following two cases: false detections (framed with a red line) and true detections that would not be detected by our content-based version of the method (framed with a green line).

to integrate content- and context-based face detectors. The approach is based on learning, in a discriminative way, the patterns of window activations that best represent true positive detections. The resulting scoring procedure behaves in a consistent way and allows a very good ranking of detections. The method has been tested in a set of challenging databases. These experiments show that context-based information plays an important role for getting good results. It has also been compared to some of the best published methods using public databases. In all cases, our method ranks the first.

The proposed methodology is not specific for face detection. It can be easily extended to related problems (i.e. detection of car wheels or bicycle seats) in a straightforward way because it does not make any assumption about the nature of the single detectors that are integrated.

The method has been defined in a pure bottom-up way, but as a future line of research, we are considering the addition of a top-down component. This component could be interesting for increasing the evidence of "*faceness*" in large scale faces by the use of a more complex face model. It is interesting to point out that all content-based methods rely on very low resolution face models (20×20 pixels) to take a decision. This limitation, that is imposed by efficiency constraints, could be relaxed by this component with a very low efficiency cost. Moreover, a top-down component could also make use of more specialized object detectors (such as eyes, mouth, shoulders, hair, etc.) which could be directly integrated in the current approach.

ACKNOWLEDGEMENTS

This work has been partially supported by MICINN Grants TIN2009-14404-C02 and CONSOLIDER-INGENIO 2010 (CSD2007-00018)

REFERENCES

Atanasoaei, C., McCool, C., and Marcel, S. (2010). A principled approach to remove false alarms by modelling the context of a face detector. In *Proc. BMVC*, pages 17.1–17.11. BMVA Press.

Bourdev, L. and Brandt, J. (2005). Robust object detection via soft cascade. In *Proceedings of IEEE Conference on CVPR'05 - Volume 2*, pages 236–243, Washington, USA.

Bourdev, L. and Malik, J. (2009). Poselets: Body part detectors trained using 3d human pose annotations. In *International Conference on Computer Vision*.

Everingham, M., Van Gool, L., and et. al. The PASCAL Visual Object Classes Challenge (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html.

Felzenszwalb, P. F., Girshick, R. B., and et. al (2010). Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 32:1627–1645.

Frome, A., Cheung, G., and et. al (2009). Large-scale privacy protection in google street view. In *ICCV*, pages 2373–2380.

H. Takatsuka, M. T. and Okutomi, M. (2007). Distribution-based face detection using calibrated boosted cascade classifier. In *Proc. ICIAP*, pages 351–356.

Jain, V. and Learned-Miller, E. (2010). Fddb: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst.

Jain, V. and Learned-Miller, E. (2011). Online Domain-Adaptation of a Pre-Trained Cascade of Classifiers. In *CVPR*.

Kienzle, W., Bakir, G., Franz, M., and Schölkopf, B. (2004). Face detection - efficient and rank deficient. In *NIPS*.

Kruppa, H. and Schiele, B. (2003). Using local context to improve face detection. In *British Machine Vision Conference (BMVC'03)*, Norwich, UK.

Lin, H.-T., Lin, C.-J., and Weng, R. C. (2007). A note on Platt's probabilistic outputs for support vector machines. *Mach. Learn.*, 68:267–276.

Maji, S. and Malik, J. (2009). Object detection using a max-margin hough transform. In *Proc. CVPR*. IEEE.

Mikolajczyk, K., Schmid, C., and Zisserman, A. (2004). Human detection based on a probabilistic assembly of robust part detectors. In *European Conference on Computer Vision*. Springer-Verlag.

Orabona, F., Keshet, J., and Caputo, B. (2009). Bounded kernel-based online learning. *J. Mach. Learn. Res.*, 10:2643–2666.

Rowley, H. A., Baluja, S., and Kanade, T. (1995). Human face detection in visual scenes. In Touretzky, D. S., Mozer, M., and Hasselmo, M. E., editors, *NIPS*, pages 875–881. MIT Press.

Subburaman, V. B. and Marcel, S. (2010). Fast bounding box estimation based face detection. In *ECCV, Workshop on Face Detection: Where we are, and what next?*

Viola, P. A. and Jones, M. J. (2001). Rapid object detection using a boosted cascade of simple features. In *CVPR (1)*, pages 511–518. IEEE Computer Society.

Xiao, R., Zhu, H., Sun, H., and Tang, X. (2007). Dynamic cascades for face detection. *Computer Vision, IEEE International Conference on*, 0:1–8.

Yang, M.-H., Kriegman, D. J., and Ahuja, N. (2002). Detecting faces in images: A survey. *IEEE TPAMI*, 24(1):34–58.

Zhang, C. and Zhang, Z. (2010). A survey of recent advances in face detection. In *Microsoft Res. Tech. Rep., MSR-TR-2010-66*.