Qualitative Portrait Classification

Georgia Albuquerque, Timo Stich, Marcus Magnor

Computer Graphics Lab, TU Braunschweig
Mühlenpfordtstr. 23, 38106 Braunschweig, Germany
Email: {georgia,stich,magnor}@cg.tu-bs.de

Abstract

Due to recent advances in high-quality digital photography, taking a large series of images is very inexpensive. Especially in portrait situations, this results in a possible advantage because subjects often feel uncomfortable during acquisition. Selecting from a larger set of images increases the chance of a more satisfying outcome. However, the selection process is not easy and time consuming as only a small number of images is typically considered as aesthetically pleasing. In this work, we propose a machine learning approach to mimic the selection process of a human subject. After a short training period, a large set of images can be classified instantly into two categories, *good* or *bad*. With the proposed automatic pre-selection, the advantage of digital photography for portrait images is brought to a new level.

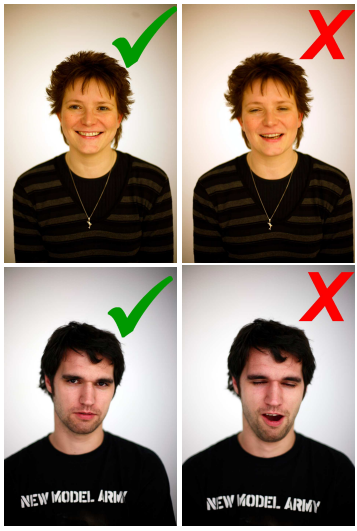

Figure 1: Samples of *good* and *bad* portraits.

1 Introduction

In the early days of photography, it was common to have people waiting for several minutes in front of the camera until the image could impress sufficiently the photographic film. It was common to retouch some areas of the photo, such as eyes, mouth and hands, which could hardly be kept immobile for long time. Later, the technology of fast films allowed to overcome this problem, but yet at a high cost of photographic material, such as special pellicles and a complex film development process.

Recent advances in high-quality digital photography allow the acquisition of large series of images in a very inexpensive way. Modern digital cameras are capable of taking and storing hundreds of high-resolution pictures within seconds. Some of them allow capturing video streams which are converted later to a sequence of individual photos. These improvements turn out in a low cost per picture for the photographer and more comfort for the photographed subject. Especially in portrait situations, such as studio or interview photos, the subject may be entertained or involved in a conversation while a fast sequence of pictures is captured. Later, the best photos from the sequence may be extracted, and the others may be discarded or digitally processed. In general, photos obtained in that way look much more natural and expressive because the subject is not worried about positioning himself into an appropriate pose.

Selecting from a large set of photos increases the chance for a better outcome. However, the selection process is very time consuming and usually not easy. While the price for acquiring and storing the images drops, the amount of produced pictures makes it tedious to run the selection phase manually. In other words, a non-automated classification

of several hundred photos quickly becomes an overwhelming task. The main reason is that selection should be applied involving complex criteria over a large set, where only some few elements are considered aesthetically pleasing. Moreover, the group of chosen photos may be highly dependent on the person running the selection, even if the acceptance rules are clearly stated.

For these reasons an automated solution for classifying portraits becomes very interesting. A semi-automated portrait classifier should be able to get rid of portraits that are obviously bad according to some criteria. The remaining portraits may be then finely classified by a human. A completely automated classifier system should be able to indicate securely the photos in the set which would most fit the predilection of the system user. For both cases, it is necessary to mimic the selection process of a human user. This is a non-trivial task, because the rules for selection of pleasant and aesthetic portraits are very problematic to be expressed in objective ways.

1.1 Related Work

Image classification is an important part of remote sensing, image analysis, and pattern recognition. It has a wide range of applications in many different areas like, for example, classification of satellite images and air photos [8], image content based search engines in the Web[12], and Biometric recognition[4][6]. Although these works present particularities dependent on the modality, all of them classify images into classes defined by previously known subjects.

The main challenges in this work are likewise how to identify facial elements, extract features and the final classification. Moreover the system should be able to learn the user's taste, based on user input.

Considering the actual research work on recognizing and classifying subject portraits, the *Image Intelligence* [2] technology from *Fujifilm* brings some approaches for face picture classification based on face recognition. On the World Wide Web, a good example of portrait a recognition-classification system is the Riya System [12]. It is a new kind of visual search engine specialized in facial recognition cataloguing. It offers the users the possibility to find similar faces and objects on many images across the web.

Most research made with facial picture classification considers face recognition. In other words, the main objective is to recognise *who* is the subject in the picture. That is the goal, for example, in biometrics authentication area. Although we use some common techniques, we are mainly interested in classifying pictures into qualitative categories such as good or bad.

In this point of view, some similar work can be found in the area of affective computing, an application of pattern recognition introduced by Picard [11]. The recognition of facial expressions, brings many contributions to the face classification area. Pantic and Rothkrantz[9] provide an overview over the area of facial expression analysis.

Some contributions for the qualitative classification of portraits were introduced in [13] with their concept of identifying neutral faces. The main motivation for this work is augmenting the accuracy of an arbitrary authentication algorithm by feeding it with a neutral face. Moreover, an impressive method for portrait images processing is presented in [7]. In contrast to a pattern recognition approach, the method automatically increases the predicted attractiveness rating of the face image. However, their method actually distorts the face and the beautification comes at the cost of identity loss.

All the cited work have some indirect relation to our topic, but as far as we know, no work was developed in order to separate portraits in adequation qualitative classes, as *good* or *bad* (Figure 1).

2 Portrait Classification

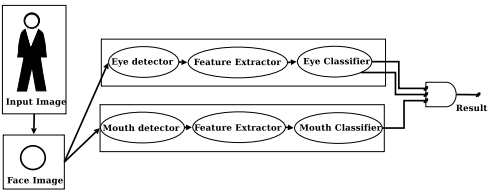

Figure 2: First the face is selected, then eyes and mouth regions are separated and used to feed the feature extraction module. Finally, the features are classified. If both eyes and mouth are classified as *good* the picture is considered good, else it is considered *bad*.

The main goal of this work is, using machine

learn techniques, mimic the selection process of a human subject allowing later an automatic selection of good portrait photos from a sequence of natural poses. This section presents an overview of our system with the main stages needed to evaluate a portrait image and classify it as a *good* or *bad* shot (Figure 2).

As input, our method receives pictures containing a single person photographed in a frontal pose. It inspects them to determine whether the presented picture is a good shot or not. However, the perception of how "good" a picture is very subjective. For example, some persons may prefer portraits where the photographed subject appears smiling, while others may consider the smile a sign of frivolity or distrust. This personal aspect restrains the acquisition of an enough general training set for the application. We propose two distinct approaches to overcome this problem. The difference between these approaches depends only on the origin of the training set.

Since mouth and eyes play extraordinary roles in facial expressions, we limit the classification to those features. In this way, we can focus on smaller, more significant regions of the picture.

The first approach is a more general system, which tries to learn the common sense of pleasantness of an image. The idea is to identify and discard face images with undesired details as, for example, closed eyes or mouths in movement (often present when someone speaks). For this mode of operation, we trained our portrait classifier with samples of eyes and mouth images in many different configurations. The samples were generated from images of a face database and were labelled as *good* or *bad*, according to simple criteria as discussed previously.

The second approach requires additionally two input image sets selected by the user, one containing good shots and the other disliked pictures. Our system then tries to learn the person's taste for the following classifications.

2.1 Detecting Face Regions

At any classification procedure, the feature extraction process plays a very important role. Largely applied in the field of machine learning and pattern recognition, feature extraction is an intelligent way to reduce the dimensionality of a large set of data, in our case, face images. The reduction of the dimensionality minimizes the amount of resources required to describe an image, and thus the resource demand for the training algorithms. We have chosen Adaptative Boosting (AdaBoost)[3] combined with Principal Component Analyses (PCA)[10] in order to define our feature set. The AdaBoost module detects the eyes and mouth regions, while the PCA module extracts the main components of these regions.

Adaptative Boosting, also called AdaBoost, was introduced in 1995 in [3]. It is a special case of Boosting, which is a general way to increase the accuracy of any given learning algorithm. We employ for our face detection module a very interesting AdaBoost implementation proposed by Viola and Jones in [16] to select the eyes and mouth regions of a portrait.

For our approach we trained two different AdaBoost modules, one for eye detection and another for mouth detection. Both classifiers are trained by two image sets. A positive set, containing images from the object of interest and a negative set, containing background images i.e. any possible image that does not correspond to the object.

To train the cascade of classifiers, the AdaBoost algorithm tries to meet an adequate trade-off between the quantity of features chosen for the classifier and the time necessary to compute the classifier. Each stage in the cascade reduces the false positive rate and decreases the detection rate as well. Each classifier stage is trained by adding features until a desired level of detection and false positive rates for the respective stage is reached. Similarly, stages are added until the overall desired levels of detection and false positive rates are met (see Figure 3 and 4).

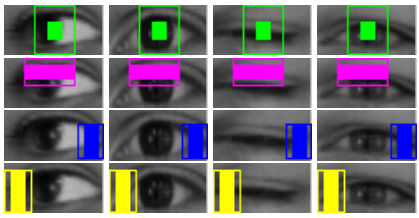

Figure 3: Haar-like features for the first stage of the AdaBoost algorithm trained to detect eyes. The first up to fourth features are marked in green, magenta, blue, and yellow, respectively.

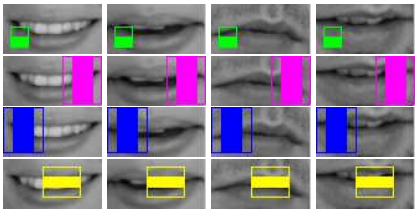

Figure 4: Haar-like features for the first stage of the AdaBoost algorithm trained to detect mouths. First up to fourth features are marked in green, magenta, blue, and yellow, respectively.

2.2 Feature Extraction

After detecting the regions of interest (eyes and mouth), the next step is to extract specific features of these regions. These are usually more robust than working directly on pixel values and can additionally encode ad-hoc domain knowledge which is otherwise difficult to learn using a finite quantity of training data. Using Principal Component Analysis (PCA) [14] is a common approach to find such robust feature descriptions in images. We follow this approach, by using the set of eyes and mouths previously selected by the detection module to calculate an eigenobject basis. Another possibility for a training set, used for the general classification mode, are the images used to train the detector module. This approach has the advantage that the images are better controlled, giving better results.

Figure 5 shows the ten most significant eigenobjects calculated for the eye and mouth spaces respectively. Interesting aspects from the eigenobject pictures can be seen. For example, the first eigenobject picture of the eye space represents eyes that look straight very well, while the second an third eigenobjects represent eyes looking to the right and left, respectively. In the case of the mouth space, the first eigenobject represents a laughing mouth, while the second represents a closed mouth. The appearance of the eigenobjects supports the hypothesis that our training set is very well represented in terms of the extracted eigenobjects.

2.3 Classification

The last step of our method is the final classification which partitions the input into a number of categories or classes[5]. In the case of recognizing whether a face image is a good shot or not, two final classes are defined: one class for the ”good” and another for the ”bad” pictures.

Figure 5: 5 first calculated eigenobjects for the eye and mouth space, respectively. Sorted from the more significant to the less significant.

The classification approach adopted in this work follows a geometric approach based on decision boundaries. We chose to use a Support Vector Machine[15] (SVM) classifier in order to categorize the final selected features. The introduction of support vector classifiers by Vapnik[15] is one of the most interesting advances in classifier design. The SVM maximizes the margin between the classes by selecting a minimum number of support vectors. The SVM algorithm is now-a-days one of the most commonly used classifiers and has many advantages, e.g., it can generate nonlinear classifiers with a very good generalization performance, even when using a small training set. Furthermore, when a large training set is used, the SVM classifier is able to select the minimal set of support vectors. This minimizes the computing requirements when testing new samples and avoids overfitting.

For our approach we trained two distinct classification modules, one for eye and another for mouth classification. If both eyes and mouth are classified as *good* the picture is considered good, else it is considered *bad*. The training sets used for the classification modules are explained in detail in the sections 3.3.1 and 3.3.2.

3 Results

This section describes some results obtained using our approach to classify portraits, including details about the structure and training of the different modules presented in Section 2.

In order to demonstrate the efficiency of our approach for portrait classification, we performed experiments with a pre-stored sequence of face images from different subjects in different poses. The used image database contains both training and test sets, but the composition of such sets variate depending

on the experiment.

3.1 Face Database

The first step in the implementation of the system was the creation of a face database, that is a database of face portraits. The major incentive to create a new database comes from the need of a large quantity of data and the idea of training and testing the system with real, i.e. not pre-processed images. The images were taken in a semi-controlled environment: Only frontal face pictures avoiding head rotations of more than 15 degrees and each picture including only one person. Moreover, the image acquisition was made at different days, and though under different conditions of lighting.

The camera utilized during the acquisition of the database images was a Canon EOS 5D, with shutter speed of $1/6$ Sec., aperture of F1.6, Lens 50mm and Focal Length 50.0mm. Furthermore, the shots were taken in a *Continuous Shot Mode* of one picture per second and stored in a Canon proprietary raw format to avoid compression artefacts.

The database contains a total of 1262 images from 12 distinct subjects at different facial poses, as for example: Open and closed eyes, looking toward different directions, laughing, speaking, yawning, etc.

3.2 Object Selection

We implemented two object selection modules: One for eye detection and another for mouth detection of face images. Both modules are similar. They differ only in the image training set and some simple restrictions discussed in section 3.2.1

For the eye detection module, we trained the system using 834 positive samples in different configurations and 2168 negative samples containing pieces of the background of some face images. After the training process, a classifier with 16 stages was achieved.

The mouth detection module was trained using 472 positive samples and 1884 negative samples. Figures 6 and 7 show some examples of positive and negative samples used to train the detector for eyes and mouths, respectively.

The training time for the AdaBoost algorithm is rather long. It takes several hours to accomplish a satisfying object detector. However, once the detector is adequately trained, the features that compose the classifier for each AdaBoost stage can be stored. After the training, the classifiers detect objects in real-time.

Figure 6: Example of images from the eye and mouth databases, to train the eye and mouth detection modules, respectively.

Figure 7: Examples of negative samples to train the eye and mouth detection modules, respectively.

3.2.1 Detection Results

The face selection module extracts the face from an input image and feeds it into the selection modules for eyes and mouths. It was implemented using a stage classifier with 24 stages, pre-defined in the face detection application from OpenCV[1]. The detector returns any possible occurrence of a face in a given image. Occasionally, the stage classifier misclassifies one region and extracts some non-face images. We call such regions a false positive detection. In general, the face selector module achieved a high detection rate. About 99% of the faces were detected successfully in a test set containing 645 pictures from 10 different subjects. False positive detections were extracted in 12% of the pictures additionally to the correct face. However, the false positive detections are not meaningful for the face detection module. Even if the program extracts some other non-face regions, the application searches in each of the candidate regions for at least two eyes and one mouth occurrences. Because the eye and mouth detection modules are unlikely to find target objects in a non-face region, this kind of region is naturally discarded.

In contrast to the face detection case, false positive detections arising during the eye and mouth selection may be problematic. That is because no filtering is made after the regions are selected.

Thus, we apply some simple location constraints for the eye and mouth detector in order to reduce the false positive detection. Our eye detection module achieved a detection rate of 90% over 332 previously selected faces, i.e. 664 eyes. Additional false positive detections were found in 7% of the images. Such set of portraits considered 5 distinct subjects, where three of them did not contributed to the training set of the eye detector. Considering the 10% of non-detected eyes, about 75% of them consisted of closed eyes. In this case, it is not a problem to reject the picture, since a portrait with closed eyes is not considered a *good* portrait. Table 1 shows the individual results per subject. The second column indicates the rate of eyes correctly extracted from the quantity in column 4. The false positive indicator corresponds to the quota of portraits where missdetected regions were additionally found. Similarly for the mouth selection, the system achieved 97% of successful detection over the same set of portraits. Table 2 shows the individual results per subject.

Figure 8 shows an example of the complete object detection process, including the face and face components selection.

Subject	Detection Rate	False Positives	Eye Quantity
Person 1	91%	13%	150
Person 2	80%	5%	110
Person 3	93%	2%	110
Person 4	97%	9%	154
Person 5	86%	24%	140

Table 1: Eye detection, individual results. The pictures of the last three persons in the table did not contribute to the training set for the eye detection module.

Subject	Detection Rate	False Positives	Mouth Quantity
Person 1	97%	19%	70
Person 2	96%	36%	55
Person 3	96%	35%	55
Person 4	97%	66%	77
Person 5	97%	47%	70

Table 2: Mouth detection, individual results. The pictures of the last three persons in the table did not contribute to the training set for the mouth detection module.

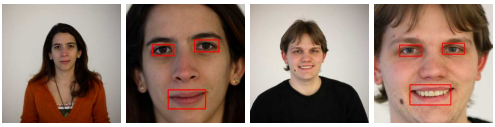

Figure 8: complete object detection process, including the face and face components selection.

3.3 Classification

This section shows some results accomplished with both classification approaches: The general approach, where the classifier is trained with previously defined sample sets of eyes and mouths labelled as *good* or *bad*. And the personal approach, where the user supplies a training face image set, representing his personal taste, labelled in the same way.

3.3.1 General Classifier

In order to build the eyes and mouth training set for the general classifier, we selected eyes and mouths from the detection module positive sample set. The labelling criteria for eyes and mouths follows a simple rule: An eye is labelled as *bad* when it is closed or looks to the right or left. Otherwise, it is labelled as *good*. A mouth is labelled as *bad* when it is open, when it makes some movement to speak, or when the smile is too large. Figures 9 and 10 show, respectively, examples of our training sets for eyes and mouths.

(a) Eyes labeled as good.

(b) Eyes labeled as bad.

Figure 9: Example of images (eyes) from the training set used for the general classification mode. Manually labeled as *good* or *bad*.

We tested the general classifier approach with 7 different subjects. Of the 7 persons used in this test, only two of them have taken part in the training set for the classification. This image test set is distinct

(a) Mouths labeled as good.

(b) Mouths labeled as bad.

Figure 10: Example of images (mouth) from the training set used for the general classification mode. Manually labeled as *good* or *bad*.

from image training set considered for the detection module and for the classifier. Table 3, shows some results for the general classification test. The column *Hit Rate* lists the percentage of face images which were correctly classified. It variates in range from 67% to 96%.

Most classification errors occurred when the images were to be classified as *bad* but were classified as *good*. Similarly to the terminology adopted in the detection phase, we call such misclassified elements a false positive classification. In the same way, we call it a false negative classification if the element was classified as *bad*, but was actually to be classified as *good*.

One can notice the strong incidence of false positive classification when classifying mouth elements. Analyzing the images where this kind of error occurs, we perceived that a considerable part of them have a mouth pose that should be considered *bad*. For such cases, our mouth classifier was not able to generalize as well as the eye classifier for dealing with counter-examples. That is explained by the fact that the mouth negative training set was not large enough. It implies that some new and undesired mouth configurations were not covered adequately by the training set. For the classification of eye elements the test rejected all occurrences of closed eyes and most of occurrences of eyes looking to right of left.

The variation in the classification hit rate for the different subjects lies mainly in the fact that between the bad images, some have more pictures with bad eyes while others with bad mouth poses. For example, most of the misclassified bad pictures for *person 5* were cases in which the person slightly pressed the lips, while the eyes were considered good. In this case, additional *bad* mouth poses are required for the training of the classifier.

	Hit Rate	**Miss rate**		**Samples**	
		FP	FN	Good	Bad
1	88%	3%	9%	20	13
2	79%	0%	21%	37	29
3	96%	0%	4%	13	14
4	79%	0%	21%	27	29
5	67%	0%	33%	32	22
6	80%	5%	15%	57	17
7	93%	7%	0%	13	16

Table 3: Classification results for the general mode, where the classifier is trained with previously defined sample sets of eyes and mouths (FP means False Positive and FN, False Negative). The persons 2 and 4 have taken participation in the training of the classifier.

3.3.2 Personal Classifier

In order to validate the personal classification mode, we executed a leave-one-out cross-validation with the same subjects from the general classification approach. For each subject, training sets are supplied, one containing face images labelled as *good* and another labelled as *bad*.

For each iteration of the test, the eyes and mouth are extracted from the pictures that compose the training set. If the picture is labelled as *good*, the extracted regions are also considered good, else the extracted regions are considered bad. After all the eye and mouth elements are extracted and labelled, the software extracts the features of the regions and uses them to train the eyes and mouth classifier, respectively. When the classifier is trained, the test picture can be classified. In a similar process to the used in the training phase, the eye and mouth elements from the test image are detected and its features are then evaluated by corresponding classifiers. If the two detected eyes and the mouth are classified as *good*, the picture is, as well, indicated to be good. Otherwise it is considered a bad picture.

The results of the leave-one-out cross-validation are shown in table 4. The column *Hit rate* indicates the percentage of the pictures that were classified accordingly with the original labels. For misclassified portraits, we indicate also the rate of false positive and false negative classification. The last two columns depict the amount of pictures for each subject and its original distribution in good and bad sets. More classification results can be seen in the Figure 12.

	Hit rate	Miss rate		Samples	
		FP	FN	Good	Bad
1	78%	19%	3%	20	12
2	70%	12%	18%	31	35
3	89%	7%	4%	13	14
4	89%	5,5%	5,5%	27	29
5	76%	7%	17%	32	22
6	72%	20%	8%	56	18
7	74%	16%	0%	13	17

Table 4: Results of the leave-one-out cross-validation for the personal classifier (FP means False Positive and FN, False Negative). Invidual results for 7 subjects

4 Conclusion & Future Work

In this work, we proposed a procedure to efficiently carry out a qualitative classification of portraits in a machine learning environment. To the best of our knowledge, we are the first to establish a workflow that entails such task. Our workflow brings together established techniques used in the area of feature extraction and machine learning. We are able to classify portraits within qualitative categories, such as good/pleasant and bad/unpleasant shots, with an accuracy rate up to 96 percent.

Our system is also capable of learning the user's preferences or taste. The user must not know details about the underlying classification system, but he indicates his taste by selecting example images for good and bad photos. We consider these aspects fundamental for the application of an qualitative image classification in realistic applications. Although our approach deals only with eyes and mouth regions of a portrait when inferring its quality the performance of our classifier was very good in general. To further increase the performance, the system may be extended to take in account other portrait elements, such as eyebrows, ears or the nose. Since the single detection units are independent, such additional features are easily added.

Another very interesting extension of our application would be to select the best shots from video sequences. The overhead to implement this extension is minimal because the set of video frames corresponds directly to a set of portraits in the face database.

References

[1] Intel Open Source Computer Vision Library. Website. http://www.intel.com/research/mrl/research/opencv/.

[2] FUJIFILM Corporation. Image intelligence. Website. http://www.fujifilm.com/image_intelligence/.

[3] Yoav Freunf and Robert E. Schapire. A decision-theoretic generalization of on-line learning and application to boosting. *In Computational Learning Theory: Eurocolt' 95*, pages 23–37, 1995.

[4] Anil K. Jain, Ruud Bolle, and Sharath Pankanti. *Biometrics: Personal Identification in Networked Society*. Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[5] Anil K. Jain, Robert P. W. Duin, and Jianchang Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[6] S. Y. Kung, M. W. Mak, and S. H. Lin. *Biometric Authentication: A Machine Learning Approach*. Prentice Hall, 2004.

[7] Tommer Leyvand, Daniel Cohen-Or, Gideon Dror, and Dani Lischinski. Digital face beautification. In *SIGGRAPH '06: ACM SIGGRAPH 2006 Sketches*, page 169, New York, NY, USA, 2006. ACM Press.

[8] Duda T.; Canty M. Unsupervised classification of satellite imagery: choosing a good algorithm. *International Journal of Remote Sensing*, 23:2193–2212, 2002.

[9] Maja Pantic and Leon J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.

[10] K. Pearson. On lines and planes of closest fit to systems of points in space. *London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. Sixth Series.

[11] Rosalind W. Picard. *Affective Computing*. MIT Press, 1997.

[12] Inc. Riya. Rija - visual search. Website. http://www.riya.com/.

[13] Yingli Tian and Rudolf M. Bulle. Automatic detecting neutral face for face authentication. In *AAAI-03 Spring Symposium on Intelligent Multimedia Knowledge Management*, 2003.

[14] Matthew Turk and Alex Paul Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[15] V. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, 1998.

[16] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *In IEEE Conference on Computer Vision and Pattern Recognition*, pages 609–615, 2001.

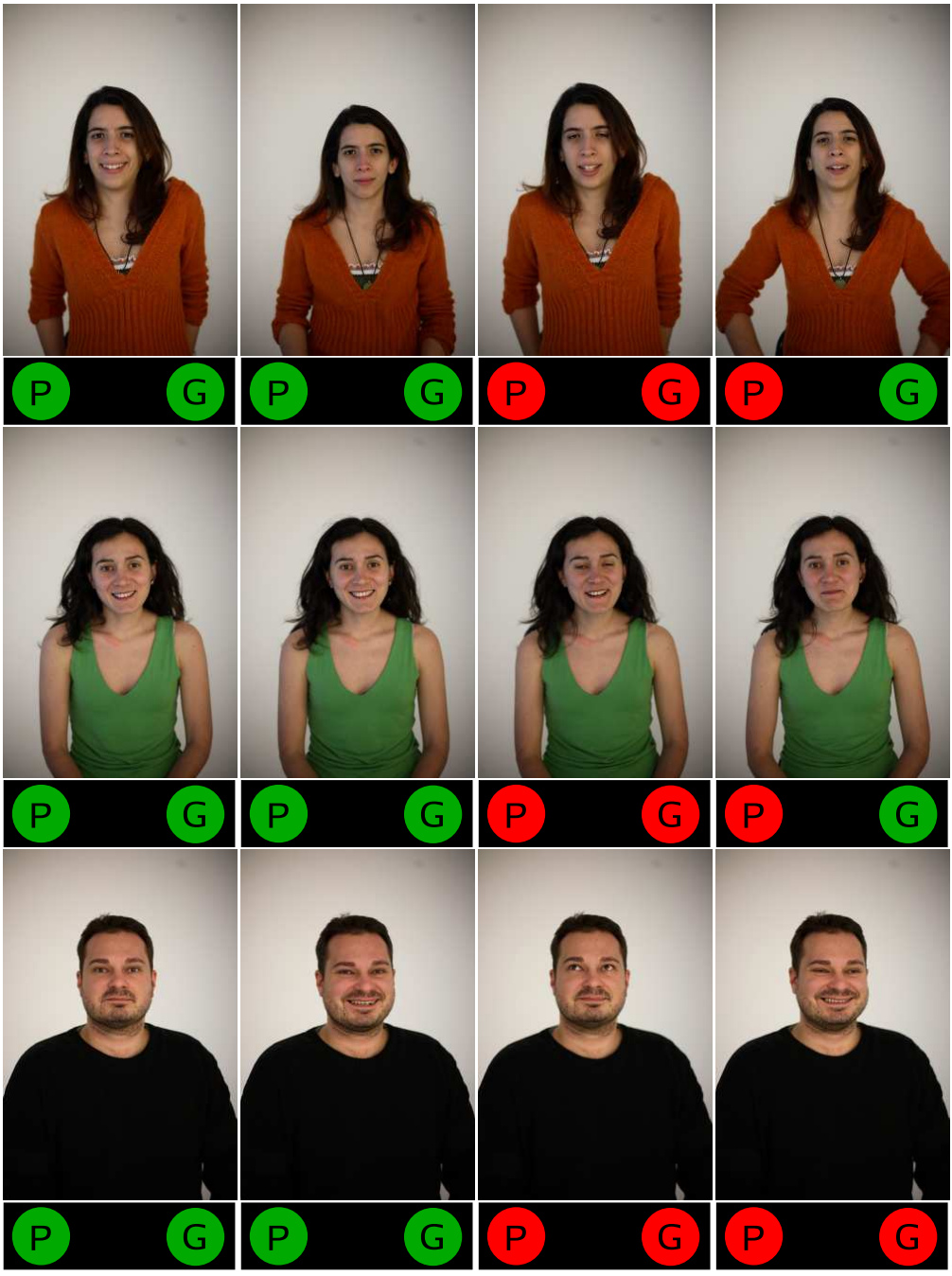

Figure 11: Example classfication results. The first circle is the vote of the Personal Classifier trained with the taste of one of the authors. The second shows the vote of the General Classifier.(green for good, red for bad)

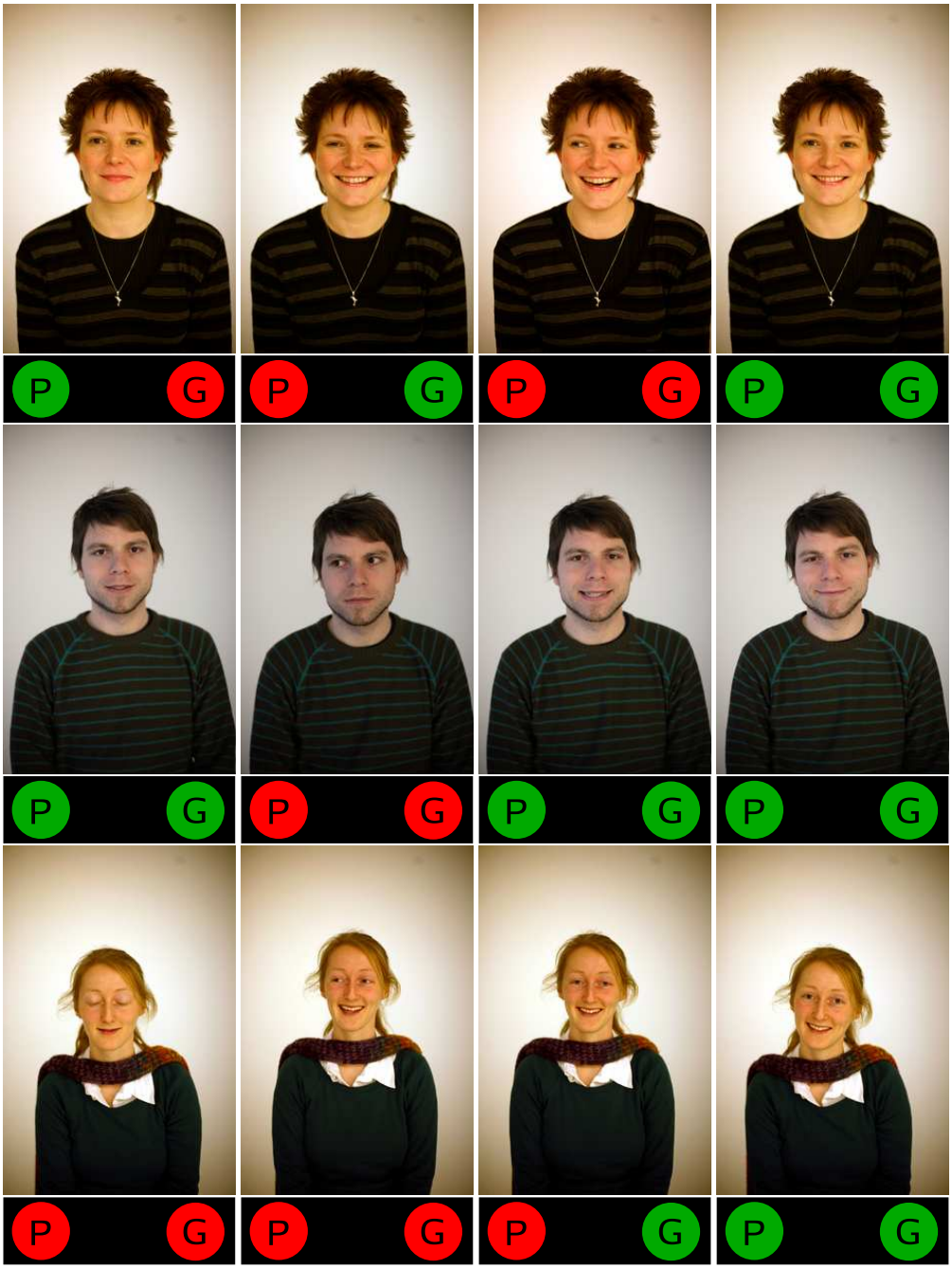

Figure 12: Example classfication results. The first is the vote of the Personal Classifier trained with the taste of one of the authors. The second shows the vote of the General Classifier. (green for good, red for bad)