Benchmarking Face Detection in a Mobile/Tablet Environment

Sébastien Marcel, Cosmin Atanasoaei and Chris McCool
Idiap Research Institute
Centre du Parc - rue Marconi 19
CH-1920 Martigny, Suisse
{sebastien.marcel}@idiap.ch

1. Introduction

Despite recent major advances, face detection is still a challenge in difficult conditions, when the face is not frontal and/or the illumination conditions are adverse. This is particularly the case when the capture device is portable such as mobile phones and now tablets. Recent mobile devices and tablets are equipped with video camera and now provide sufficient computing and memory resources to accommodate intensive processing such as the one involved in face detection, alignment and recognition. As a consequence, many new applications based on these technologies will emerged and it becomes critical to understand the limit of current technologies under such a particular mobile environment.

For this first challenge, we propose to focus only on the face detection task. Face detection is a challenging problem because faces highly vary in size, shape, color, texture and location. Their overall appearance can also be influenced by lighting conditions, facial expression, occlusion or facial features, such as beards, mustaches and glasses. Another challenging problem comes from the orientation (upright, rotated) and the pose (frontal to profile) of the face. The goal of face detection is to determine whether or not there are any faces in the image and, if present, their location. It is the crucial first step of any application that involves face processing systems. Thus, accurate and fast human face detection is the key to a successful operation.

This face detection task will be evaluated on a recent database, called MOBIO (http://www.idiap.ch/dataset/mobio) that is one of the main achievement of the EU FP7 project MOBIO (http://www.mobioproject.org).

The MOBIO database was captured in six different sites across Europe. It consists of over 61 hours of audio-visual data from English speakers (both native and non-native) and was captured using mobile devices with limited control on the capture conditions. The database was collected over a year and a half and consists of 150 participants each with 12 sessions captured on a mobile phone and one session on a laptop. The data was captured to reflect the potential real-world challenges that could be faced when performing face processing on a mobile device such as uncontrolled quality of the video due to movement of both the subject and mobile device.

2. Data

The MOBIO database is a multi-modal database that was captured to address several issues in face, speaker and bi-modal authentication. These include model adaptation and the effect of typical degradations that can be expected when biometric traits are acquired using a hand held platform. For such studies we need:

- consistent data captured over a long period of time to study the problem of model adaptation,
- video capturing a range of variability in illumination and pose representative of realistic settings, and
- audio captured on a mobile platform with varying degrees of realistic noise.

The database consists of over 61 hours of audio-visual data (28800 videos) and was captured at six different sites in five different countries: Switzerland (IDIAP), France (LIA), Finland (UOULU), United Kingdom (UMAN and UNIS) and the Czech Republic (BUT)[1]. The data was captured exclusively on mobile devices in real-world conditions for face and speaker authentication.

This bi-modal database contains 150 participants with a female to male ratio of nearly 1:2 (51 female subjects and 99 male subjects). It was acquired in two phases. During the first phase (Phase I), from August 2008 to March 2009, six sessions separated by several weeks were recorded with the first session being a double session acquired on both the laptop and the mobile phone. The mobile phone used to capture the database was a Nokia N93i and the laptop computer was a standard 2008 MacBook. For each device the data was recorded in a high quality format. More information on the MOBIO database could be obtained from [2].

Additionally, one image was extracted from each video and was manually annotated with eye locations. This face detection task will be evaluated on these images only and not on the videos. Both the images and the annotations will be provided.

3. Protocol

The database is split into three non-overlapping partitions: one for training, one for development (or evaluation) and one for testing. The data is split so that two of the six collecting sites are used in totality for one partition. This means that the data acquisition conditions in different sites cannot be shared between any of the three partitions. This is a realistic scenario because the system designer often does not have the actual operational data in order to be able to fine tune the system parameters. The non-overlapping development and test partitions ensure that there is an unbiased performance assessment.

The training partition can be used in any way deemed appropriate. The normal use of the training set would be to train a face detector. This partition consists of the data collected from the UNIS and LIA sites and has a total of 50 subjects (13 females and 37 males).

The purpose of the development partition is to define a threshold that is then applied to the test data. The development data can also be used to estimate fusion parameters, in the case of the use of multiple face detectors, and for system optimization. This partition consists of the data collected from the UMAN and UOULU sites and has a total of 42 subjects (18 females and 24 males).

The test partition is used to provide the final performance evaluation of the system and as such should be used sparingly. No meta parameters can be learnt from this set. This partition consists of the data collected from the BUT and IDIAP sites and has a total of 58 subjects (20 females and 38 males).

4. Measure

The manual locations will be used to assess the performance of face detection and eye localization using the Jesorsky measure [1]. Indeed, from the detected face bounding box (rectangle), the position of eyes can be often inferred as the face detector has been trained from cropped face images according to the aforementioned bounding box determined by annotated eye locations.

The Jesorsky measure defines how much both eye detections vary from the ground truth relative to the distance between the eyes. Let us define the detected eye positions as l_d and r_d (left and right) and the ground truth positions as l_g and r_g. The errors in pixels of each detection are $E_L = \|l_d, l_g\|$ and $E_R = \|r_d, r_g\|$ and the distance between the eyes D. Then the Jesorsky error is defined as: $\epsilon_J = \frac{max(E_L, E_R)}{D}$, which linearly relates the maximum allowed eye distance error with the distance between the eyes.

For example for a face detection with a distance between the eyes of $D = 100$ pixels, a detection with $\epsilon_J \leq 0.10$ can have the displacement between the true eye position and the inferred position (from the detection) of at most 10 pixels. We have schematically represented in Fig. 1 the areas where a valid left and right eye detection can reside (hashed red circles) for $\epsilon_J = 0.10$ and $\epsilon_J = 0.25$.

Hence, to label a detection as positive we propose to use this Jesorsky measure with the threshold $\epsilon_J = 0.25$. Note that this is **more restrictive** than the usual condition to register a positive detection: a minimum overlap of 50% with the ground truth bounding box [3]. This implies that the reported detection rate values may be smaller than the latest reported state-of-the-art results. But we are motivated into using the Jesorsky measure because it provides a better alternative to compare the accuracy of the detections obtained with different methods.

[1] The organizations involved in the data recording were the University of Oulu (OULU), Idiap Research Institute (IDIAP), University of Avignon (LIA), University of Manchester (UMAN), University of Surrey (UNIS) and the Brno University of Technology (BUT).

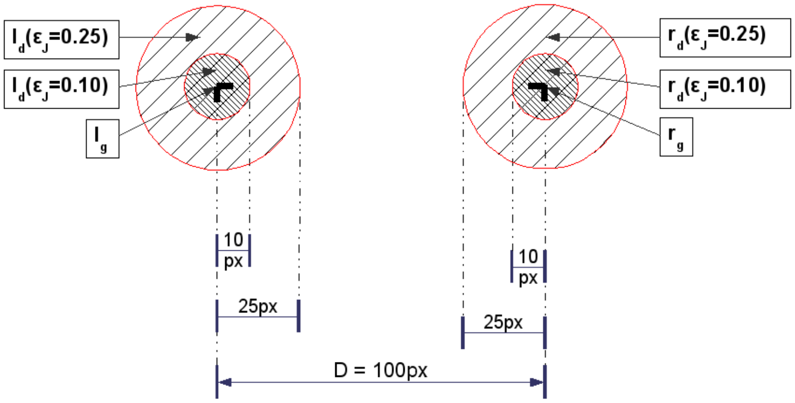
$l_d(\varepsilon_J=0.25)$
$l_d(\varepsilon_J=0.10)$
l_g
$r_d(\varepsilon_J=0.25)$
$r_d(\varepsilon_J=0.10)$
r_g
10 px
25px
10 px
25px
D = 100px

Figure 1. Areas of valid left and right eye detections (dashed) for various Jesorsky measure thresholds using the distance between the eyes as 100 pixels.

Finally, the performance of the face detection could be simply assessed by two numbers on the testing split: the True Acceptance Rate (TAR), also referred to as Detection Rate (DR), and the False Acceptance Rate (FAR). Note that several pairs of such numbers can be obtained on the development split by varying the decision threshold and hence computing a Receiver Operating Characteristic (ROC) curve plotting the DR as a function of the FAR.

5. Future work

We plan to propose more benchmarks on the MOBIO database, including eye localization, gender recognition, face recognition (both verification and identification) from still images or videos, and finally adaptive face recognition algorithms.

References

[1] O. Jesorsky, K. J. Kirchberg, and R. Frischholz. Robust face detection using the hausdorff distance. In *AVBPA '01: Proceedings of the Third International Conference on Audio- and Video-Based Biometric Person Authentication*, pages 90–95, London, UK, 2001. Springer-Verlag.

[2] C. McCool and S. Marcel. Mobio database for the icpr 2010 face and speech competition. Idiap-Com Idiap-Com-02-2009, Idiap, 11 2009.

[3] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:511–518, 2001.