Characters or Faces: A User Study on Ease of Use for HIPs

Yong Rui, Zicheng Liu, Shannon Kallin, Gavin Janke and Cem Paya

{yongrui, zliu, shannonk, gavinj, cemp}@microsoft.com

Abstract. Web-based services designed for human users are being abused by computer programs (bots). This real-world issue has recently generated a new research area called Human Interactive Proofs (HIP), whose goal is to defend services from malicious attacks by differentiating bots from human users. During the past few years, while more than a dozen HIP systems have been developed, there is little user study been done in evaluating HIP's ease of use and friendliness. In this paper, we first introduce a new HIP based on human face detection, and then report a comparative user study between this new face HIP and a more conventional character-based HIP. Study results show that the users are almost equally divided in evaluating their overall ease of use.

1 Introduction

Web services are increasingly becoming part of people's everyday life. For example, we use free email accounts to send and receive emails; we use online polls to gather people's opinion; and we use chat rooms to socialize with others. But all these Web services designed for human use are being abused by automated computer programs (bots). Malicious programmers have designed bots to register thousands of free email accounts every minute [1,3]. Bots have been used to cast votes in online polls [1]. Chat rooms and online shopping are being abused by bots as well [2, 7].

These real-world issues have recently generated a brand-new research area called Human Interactive Proofs (HIP), whose goal is to defend services from malicious attacks by differentiating bots from human users. The first idea related to HIP can be traced back to Naor who wrote an unpublished note in 1996 [7]. The first HIP system in action was developed in 1997 by researchers at Alta Vista [2]. Its goal was to prevent bots from adding URLs to the search engine to skew the search results. In recent years, more than a dozen HIP algorithms and systems have been developed, most of which are based on characters [1,3]. These character-based HIPs are main streams in today's commercial deployment, e.g., Yahoo, MSN Passport, etc. They mainly explore the gap between human and bots in terms of reading poorly printed or manipulated characters. Figure 1 shows a character HIP used in MSN Passport, which consists of distorted characters and random arcs. A user needs to recognize the characters and correctly types in the space below the HIP to prove he/she is a human.

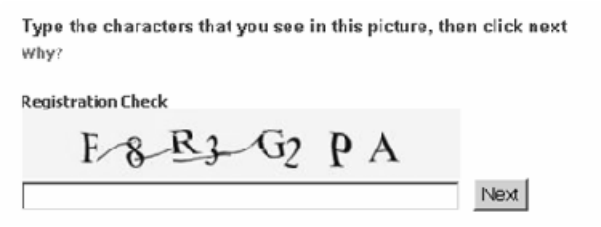


Fig. 1. An example character HIP

The MSN Passport HIP is similar to other character HIP in that it uses distorted and manipulated texts. However, it has an additional defense due to its segmentation difficulty, e.g., the arcs [4]. We will use this particular character HIP to represent the class of character HIPs in the rest of this paper.

Character HIPs are the mostly widely used HIPs in today's commercial sites, because of their ease of use, ease of implementation and universality. The "universality" property requires a HIP to be usable by people from different countries. An English-digit based audio HIP, for example, does not satisfy the universality property as people who do not understand English cannot use the HIP. Universality is especially important in practice as it eliminates the localization effort for sites such as Yahoo or MSN. (See [8] for other good HIP properties).

We recently developed a HIP, which is completely different from character HIPs, yet also satisfies the universality property. This new HIP is based on human face and facial feature detection. In fact, it is even more universal than character HIPs, as people all know human faces, regardless where they come from. On the other hand, face detection and facial feature (e.g., eyes, mouth, nose, etc.) detection have been very difficult for machines, even after decades of research. Non-frontal faces, asymmetrical faces, dim/bright lighting conditions, and cluttered background make the task even more difficult for machines, while human have no problem in those situations. In [8], we reported detailed experiments on the robustness of the face HIP to malicious attacks from the best face detectors [5,11,12] and facial feature detectors [9] available today. Results show that the face HIP is robust at a rate of 2 out of a million. For details of the algorithms and attacks, the readers are referred to [8]. In this paper will concentrate on the use-friendliness aspect of the face HIP.

The face HIP works as follows. Per each user request, it automatically synthesizes an image with a distorted face embedded in a clustered background. The user is asked to first find the face and then click on 4 points (2 eyes and 2 mouth corners) on the face. If the user can correctly identify these points, the face HIP concludes the user is a human; otherwise, the user is a machine.

During the past few years, while more than a dozen HIP systems have been developed, there is little user study been done in evaluating HIP's ease of use and friendliness. But in reality, ease of use is as important as the robustness (to attack) of a HIP. Good user experience is becoming increasingly important as HIPs are not only used in

one-time activities (e.g., registering an account), but also in *recurring* transactions (e.g., challenge-response systems against spam). In this paper, we will report a comparative user study between this face HIP and the MSN character HIP. Study results show that the users are almost equally divided between the two HIPs in terms of overall experience. The rest of the paper is organized as follows. In Section 2, we describe the face HIP. In Section 3, we discuss the user study design and methodology. We present the study results in Section 4 and give concluding remarks in Section 5.

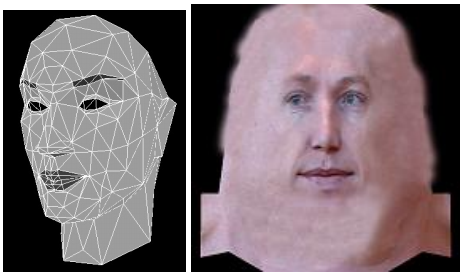
2 The Face HIP

The details on how to create the face HIP is reported in [8]. For completeness of this paper, we give a brief description of the HIP algorithm in this section.

Human faces are arguably the most familiar object to humans, rendering it possibly the best candidate for HIP. Regardless of nationalities, culture differences or educational background, we all recognize human faces. In fact, our ability is so good that we can recognize human faces even if they are distorted, partially occluded, or in bad lighting conditions.

Computer vision researchers have long been interested in developing automated face detection algorithms. A good survey paper on this topic is [10]. In general face detection algorithms can be classified into four categories: knowledge-based, feature-based, template matching, appearance-based. So far, the fourth approach is the most successful one [10].

In spite of decades of hard research on face and facial feature detection, today's best detectors still suffer from several main limitations including the assumption that **faces are symmetric**, the difficulties of handling arbitrary **head rotations**, **arbitrary lighting**, and **cluttered background**. These conditions are among the most difficult cases for automated face detection, yet we human seldom have any problem under those conditions. If we use the above 4 conditions to design a HIP test, it can take



(a) (b)

Fig. 2. (a) The 3D wire model of a generic head. (b) The cylindrical head texture map of an arbitrary person

advantage of the large detection gap between human and machine. Indeed, this gap motivates our design of the face HIP.

We next use a concrete example to illustrate how to automatically generate a face HIP test image, taking into account of the 4 conditions discussed above. For clarity, we use *F* to indicate a foreground object in an image, e.g., a face; *B* to indicate the background in an image; *I* to indicate the whole image (i.e., foreground and background); and *T* to indicate cylindrical texture map.

[**Procedure**] Generating a face HIP test image

[**Input**] The only inputs to our algorithm are the 3D wire model of a generic head (see Figure 2 (a)) and a 512 x 512 cylindrical texture map *Tm* of an arbitrary person (see Figure 2 (b)). Note that any person's texture map will work in our system and from that single texture map we can in theory generate infinite number of test images.

[**Output**] A 320 x 320 test image I_F (see Figure 5) with ground truth (i.e., face location and facial feature locations).

1. Confusion texture map *Tc* generation

   This process takes advantage of the **Cluttered Background** limitation to design the HIP test. The 512 x 512 confusion texture map *Tc* (see Figure 3) is obtained by moving facial features (e.g., eyes, nose and mouth) in Figure 2 (b) to different places such that the "face" no longer looks like a face.

2. Global head transformation

   Because we have the 3D wire model (see Figure 2 (a)), we can easily generate any global head transformations we want. Specifically, the transformations include translation, scaling, and rotation of the head. Translation controls where we want to position the head in the final image I_F. Scaling controls the size of the head, and rotation can be around all the three x, y, and z axes. At run time, we randomly select the global head transformation parameters and apply them to the 3D wire model texture-mapped with the input texture *Tm*. This process takes ad-

Fig. 3. The confusion texture map Tc, is generated by randomly moving facial features (e.g., eyes, nose and mouth) in Fig 2 (b) to different places such that the "face" no longer looks like a face

(a) (b)

Fig. 4. (a) The head after global transformation and facial feature deformation. We denote this head by F_h. (b) The confusion head after global transformation and facial feature deformation. We denote this head by Fc

vantage of the **Head Orientations** limitation to design the HIP test.

3. Local facial feature deformations
   The local facial feature deformations are used to modify the facial feature positions so that they are slightly deviated from their original positions and shapes. This deformation process takes advantage of the **Face Symmetry** limitation to design the HIP test. Each geometric deformation is represented as a vector of vertex differences. We have designed a set of geometric deformations including the vertical and horizontal translations of the left eye, right eye, left eyebrow, right eyebrow, left mouth corner, and right mouth corner. Each geometric deformation is associated with a random coefficient uniformly distribution in [-1, 1], which controls the amount of deformation to be applied. At run time, we randomly select the geometric deformation coefficients and apply them to the 3D wire model. An example of a head after Steps 2 and 3 is shown in Figure 4 (a). Note that the head has been rotated and facial features deformed.

4. Confusion texture map transformation and deformation
   In this step, we conduct exactly the same Steps 2 and 3 to the confusion texture map *Tc,* instead to *Tm.* This step generates the transformed and deformed confusion head *Fc* as shown in Figure 4 (b).

5. Stage-1 image I_1 generation
   Use the confusion texture map *Tc* as the background *B* and use F_h as the foreground to generate the 320 x 320 stage-1 image I_1 [8].

6. Stage-2 image I_2 generation
   Make *L* copies of randomly shrunk *Tc* and randomly put them into image I_1 to generate the 320 x 320 stage-2 image I_2 [8]. This process takes advantage of the **Cluttered Background** limitation to design the HIP test. Note that none of the copies should occlude the key face regions including eyes, nose and mouth.

7. Final test image I_F generation (Figure 5)
   There are three steps in this stage. First, make *M* copies of the confusion head *Fc* and randomly put them into image I_2. This step takes advantage of the Cluttered Background limitation. Note that none of the copies should occlude the key face regions including eyes, nose and mouth. Second, we now have *M+1* regions in the image, where *M* of them come from *Fc* and one from F_h. Let *Avg(m), m = 0, …, M+1,* be the average intensity of region *m.* We next re-map the intensities of

Fig. 5. An example face HIP test image

each region *m* such that *Avg(m)*'s are uniformly distributed in [0,255] across the *M+1* regions, i.e., some of the regions become darker and others become brighter. This step takes advantage of the Lighting and Shading limitation

The above 7 steps take the 4 face detection limitations into account and generate the face HIP test images that are very difficult for face detectors. In [8], we reported detailed experiments on the robustness of the face HIP to malicious attacks from the best face detectors [5,11,12] and facial feature detectors [9] available today. Results show that the face HIP is robust at a rate of 2 out of a million. In the following section, we will report another aspect of the HIP – its ease of use and friendliness.

3. User Study Design and Methodology

We recruited 200 panelists from an independent research panel. To eliminate gender difference, the panelists are 50% male and 50% female. They also have different levels of internet experience, ranging from beginner, to intermediate, to advanced. Furthermore, the panelists have diverse income levels to eliminate another potential bias factor. The panelists voluntarily participate in the user study online from their own homes. This not only ensures that they do not need to change their regular online behavior, but also ensures that they are viewing the HIP images in the settings they would be most comfortable with, e.g., monitor type and size, screen resolution, contrast and brightness, etc.

The section of user study on comparing HIPs is part of a larger-scale study that concerns with other issues in MSN Passport registration (see the study flow chart in

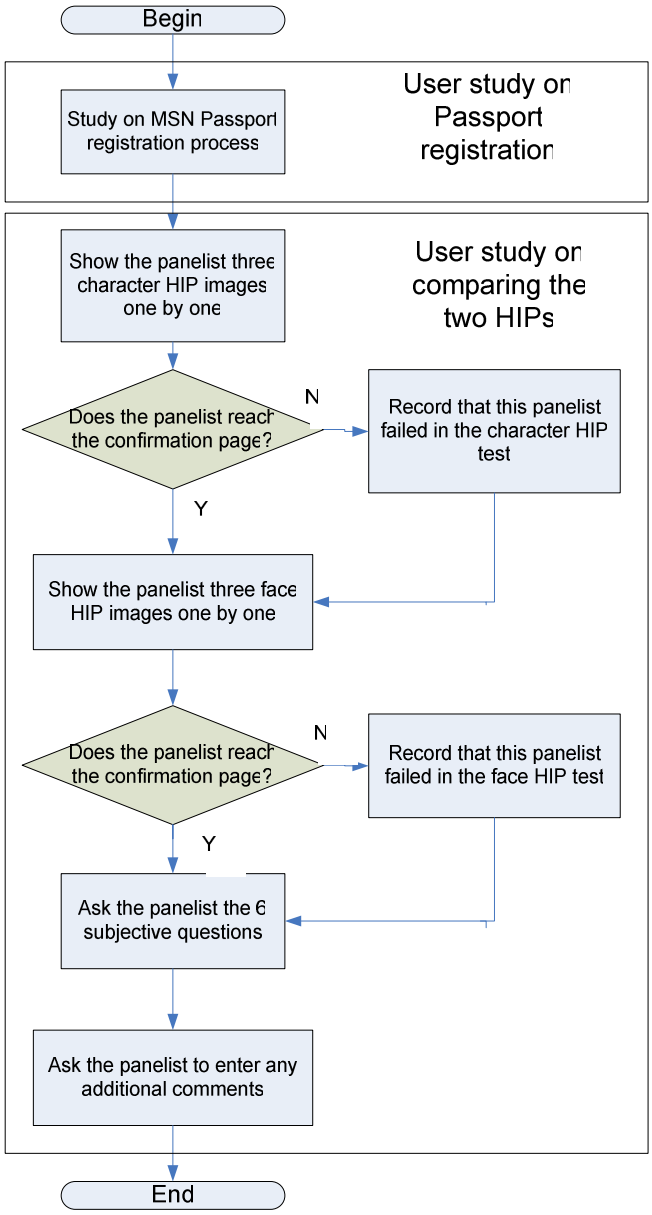


Fig. 6. . The flow chart of the user study. The first part of the study is on Passport registration generic issues, which is outside the scope of this paper. The second part of the study is on comparing the two HIPs. The first study does set up the context for the second study

Figure 6). In the study before the HIP section, the panelists go through a regular MSN

Passport registration process. As a result, the section on comparing HIPs is in full context and the panelists already know how to register in Passport and understand the purpose of putting a HIP test inside the registration process.

In the study, we use two measures to evaluate each HIP's performance, i.e., objective task performance, and subjective responses. The latter includes asking various questions to the panelists. The list of questions and panelists' responses are reported in detail in Section 4. The scenario of the objective task is at an MSN Passport registration page. The panelists are given the following instructions:

> *You may recall that at one point during the registration processes you've just evaluated, you were required to input some letters and numbers that were* ***"distorted."***
>
> *This is a* ***safety feature****, the purpose of which is to prevent automated computer programs from generating thousands of fake e-mail accounts in order to send unsolicited "spam" mail. Computers have a difficult time identifying distorted letters within an image.*
>
> *In the last 2 tasks we'd like you to take a closer look at this "character" distortion and compare it with an alternative "faces" distortion that has been developed. As you review both versions, please imagine that you are in the middle of a registration process, similar to those which you have just evaluated. You do not need to fill out the registration form, simply evaluate the images.*

Specifically, the panelists are asked to conduct the following task:

> *Without completing the registration form, please scroll down the page until you can see the image, then follow the instructions to interact with the image appropriately. Make sure that you click the "next" button to cycle through the multiple images you will be shown.*

Figure 6. The flow chart of the user study. The first part of the study is on Passport registration generic issues, which is outside the scope of this paper. The second part of the study is on comparing the two HIPs. The first study does set up the context for the second study.

If a panelist can successful pass the HIP test, by clicking on the "next" button, he/she will be presented with a similar page, but with a different HIP image. This process iterates for three (3) images. Once the panelist correctly finishes the 3rd HIP image, he/she will be greeted with a "Congratulation/Confirmation" page, indicating that he/she has finished the task. The group of 3 images can both be the character HIP and the face HIP. This "objective task" gives us an *objective* way to see if a particular HIP is easy to use – the higher the percentage of the panelists who can reach the Confirmation page, the easier the HIP is.

4. User Study Results

After finishing the *objective* task, the panelist will then be given six (6) *subjective* questions. For each question, the panelist selects a number from 1 to 7, 1 being the most "disagree" with the question, and 7 being the most "agree" with the question. For the ease of presenting results in the paper, we classify scales 1 and 2 being "Oppose", scales 3-5 being "Neutral", and scales 6-7 being "Support". In addition to the above 1-7 scale answers, we also provide panelists with a field where they can enter free-form comments (see Figure 6).

We next report the exact questions asked and the detailed results. There are two types of questions. Overall-quality questions (Q1, Q2, Q5 and Q6) evaluate the overall performance, e.g., ease of use, of a HIP. On the other hand, specific questions (Q3 and Q4) are designed to reveal more detailed findings.

4.1. Overall Findings

We classify the overall findings into two categories: "Ease of Use" and "If should Use". The results on "Ease of Use" are summarized in Table 1. The following observations can be made.

In the HIP comparison, users are split in terms of overall preference for the characters HIP and the face HIP. The levels of Ease of Use are similar for the two HIPs, not only verified by the objective task but also the subjective responses.

In fact, both Ease of Understanding Instructions **(Q1:** How difficult or easy was it to understand the instructions for interacting with the images?) and Ease of Perform-

Table 1. Overall findings: Ease of Use (* % based on responses of 6-7 on 7-point scales)

Metric	Character	Face
Objective task Success on task (reached confirmation page)	80%	78%
Ease of Understanding Instructions * **Q1:** How difficult or easy was it to understand the instructions for interacting with the images? (1 = Extremely difficult and 7 = Extremely easy)	87%	82%
Ease of Performing Task * **Q2:** How difficult or easy was it to perform this task on the "faces" version? (1 = Extremely difficult and 7 = Extremely easy)	78%	77%

Table 2. Which HIP we should use

Metric	Support	Neutral	Oppose
Character	56%	38%	7%
Face	47%	34%	19%

ing Task **(Q2:** How difficult or easy was it to perform this task on the "faces" version?) are similar for the two HIPs.

There are two questions on "If should use". For question **Q5**: Between the character HIP and the face HIP of the security feature you just reviewed, which one did you like better?, 53% prefer the face HIP and 47% prefer the character HIP – again, no significant difference between the two HIPs.

For question **Q6**: In your opinion, should Microsoft .NET Passport use the "characters"/"faces" version for users registering a Hotmail account? (1 = I would strongly oppose this and 7 = I would strongly support this), the results are listed in Table 2.

As shown in Table 2, there are 7% users who oppose character HIP. We speculate that these users either think the characters are too difficult to recognize or they don't think HIP is necessary in general. Additional research needs to be done to understand why 7% of users oppose character HIP.

It is interesting to note that there are significantly more people who oppose face HIP. From the interviews with the users, we find that some of the users who oppose face HIP think the distorted faces are offensive to them. There are another set of users who do not mind the images themselves, but they feel that the images might be offensive to other people. How to design a face HIP so that it is visually less disturbing yet difficult for bots is an interesting problem.

To summarize, although the panelists are almost equally divided on "Ease of use", they have mixed comments on "If should use" -- while more panelists (53% vs. 47%) like the face HIP, more panelists (19% vs. 7%) oppose the idea of using it. This interesting bi-modal distribution shows up again in "specific findings" in Section 4.2.1. We speculate that panelists like the face HIP because of the "seek and find then click" aspect of the task -- most panelists prefer *clicking* to *typing*. Therefore, perhaps it is the nature of the task that is liked and not the specific stimulus.

4.2. Specific Findings

4.2.1. Pleasant or not

This question is designed to reveal if distortion (of characters/faces) will pose trouble on the panelists. **Q3**: How would you rate the images of the "faces"? (1 = Very disturbing and 7 = Very pleasant)

Table 3. Pleasant or not

	Pleasant (6-7)	Neutral (3-5)	Disturbing (1-2)
Character	40%	48%	2%
Face	39%	44%	17%

17% panelist rated the face HIP as disturbing (1-2) while only 2% said the same for the character HIP. It is interesting that the panelists have a bi-modal distribution: while some commented that the faces were strange or eerie, other thought it was fun and interesting:

Eerie:

- *"It is a bit eerie to look at."*
- *"It's a little freaky looking...kinda spooky."*
- *"Don't like it; disturbing."*
- *"The faces are very creepy. The images look like severed heads."*

Fun:

- *"It seems very effective and fun for kids."*
- *"It seems a really secure and it's fun to do also."*
- *"It was interesting, and kind of cool."*
- *"I found it entertaining and useful at the same time."*

4.2.2. Size and area

We speculate that some panelists may think the areas of the HIP images maybe too big or too small. **Q4**: How would you rate the area you had to click on the image? (1 = Far too small and 7 = Far too large). The majority of panelists did not have an issue with the size / area of the characters or faces (see Table 4).

Table 4. Is the image size too large or too small

	Too large	Neutral	Too small
Character	21%	79%	1%
Face	14%	84%	3%

Fig. 7. An example where the character HIP can be difficult for human

4.2.3. Difficulties with both HIPs

For the character HIP, majority found it easy to read; however, certain letters gave them trouble when lines ran across the image (see circled area in Figure 7)

- *"The characters were easy to read, and the whole process was easy to complete."*
- *"The H looks an awful lot like the N...especially when there is a line of some sort running through."*
- *"The F's and E's can be difficult to see with the lines through them."*

For the face HIP, most users did not find it difficult to accomplish; also, some panelists mentioned that they preferred clicking on the image to typing in the characters

- *"That was surprisingly much faster and easier."*
- *"It was fast and easy. I could see the face clear."*
- *"Quicker than having to type the characters. Seems to be very easy."*
- *"This was much simpler than typing of characters."*

5. Conclusions and discussions

In this paper, we reported a comparative user study between a character HIP and a new face HIP, and have the following major findings:

- For the objective task, the panelists performed almost equally well for the two HIPs.
- For "Ease of use", the panelists rated both HIPs similarly on both the ease of performing the task and understanding the instruction of the task.
- For "If should use", while more panelists (53% vs. 47%) liked the face HIPs, there were also more panelists (19% vs. 7%) opposes the idea of using it in Passport registration page.
- There was a bi-modal distribution in panelists when asking them if the face HIP images were pleasant. While some thought the images were eerie, other thought they were fun.
- Panelists thought the size/area of both HIP images were appropriate.
- Some panelists thought the character HIPs were difficult to solve, and others prefer the face HIP (clicking) to character HIP (typing).

As the state of art on OCR technology rapidly advances, it is becoming increasingly difficult to design a character-based HIP that can be difficult for computers yet easy for humans. For example, the Gimpy HIP used at Yahoo site was broken by Mori and Malik [6], and an earlier version of MSN Passport HIP was also broken [4]. Given that face detection from images has been a difficult task for computer vision researchers for many decades, face detection and facial feature detection may be a better candidate for robust HIPs.

While the new face HIP posses many attractive features, e.g., ease of use, universality, etc., some users thought it is eerie. It will be an interesting research direction to design a HIP that has all the nice features of the current face HIP, yet is less disturbing to sensitive users.

6. Acknowledgement

We would like to thank Vividence Research for helping conduct the user study.

References

1. Ahn, L., Blum, M., and Hopper, N. J., Telling humans and computers apart (Automatically) or How lazy cryptographers do AI, Technical Report CMU-CS-02-117, February, 2002
2. AltaVista's Add URL site: altavista.com/sites/addurl/newurl
3. Baird, H.S., and Popat, K., Human Interactive Proofs and Document Image Analysis," Proc., 5th IAPR Workshop on Document Analysis Systems, Princeton, NJ, August 19-21, 2002
4. Chellapilla K., and Simard P., Using Machine Learning to Break Visual Human Interaction Proofs (HIPs), *Advances in Neural Information Processing Systems 17*, Neural Information Processing Systems (NIPS'2004), MIT Press.
5. Colmenarez A. and Huang, T. S., Face detection with information-based maximum discrimination, Proc. of IEEE CVPR, pp., 782-788, 1997
6. Mori, G. and Malik , J., Recognizing objects in adversarial clutter: breaking a visual CAPTHA, CVPR 2003, pp. I 134-141.
7. Naor, M., Verification of a human in the loop or identification via the Turing test, unpublished notes, September 13, 1996
8. Rui, Y. and Liu, Z., ARTiFACIAL: Automated Reverse Turing test using FACIAL features, ACM/Springer Multimedia Systems Journal, May 2004
9. Yan, S. C., Li, M. J., Zhang, H. J., and Cheng., Q. S., Ranking Prior Likelihoods for Bayesian Shape Localization Framework, Submitted to IEEE ICCV 2003.
10. Yang, M., Kriegman, D., and Ahuja, N., Detecting faces in images: a survey, IEEE Trans. on Pattern analysis and machine intelligence, Vol. 24, No. 1, January 2002.
11. Yang, M., Roth, D., and Ahuja, N., A SNoW-Based Face Detector, Advances in Neural Information Processing Systems 12 (NIPS 12), S.A. Solla, T.K. Leen and K.-R. Muller (eds), pp. 855--861, MIT Press, 2000.

12. Zhang, Z., Zhu, L., Li, S. and Zhang, H, Real-time multiview face detection, Proc. Int'l Conf. Automatic Face and Gesture Recognition, pp. 149-154, 2002