Robust Hand Detection

Mathias Kölsch and Matthew Turk
Department of Computer Science, University of California, Santa Barbara, CA 93106

Abstract

Vision-based hand gesture interfaces require fast and extremely robust hand detection. Here, we study view-specific hand posture detection with an object recognition method recently proposed by Viola and Jones. Training with this method is computationally very expensive, prohibiting the evaluation of many hand appearances for their suitability to detection. As one contribution of this paper, we present a frequency analysis-based method for instantaneous estimation of class separability, without the need for any training. We built detectors for the most promising candidates, their receiver operating characteristics confirming the estimates. Next, we found that classification accuracy increases with a more expressive feature type. As a third contribution, we show that further optimization of training parameters yields additional detection rate improvements. In summary, we present a systematic approach to building an extremely robust hand appearance detector, providing an important step towards easily deployable and reliable vision-based hand gesture interfaces.

1 Introduction

Vision-based interfaces (VBI) are gaining much interest recently, maybe best illustrated by the commercial success of Sony's Eye Toy, an accessory for the company's PlayStation 2: a set-top camera recognizes full-body motions and projects the player directly into the game. However, more fine-grained control, such as with hand gesture recognition, has not yet reached the same level of robustness and reliability. Outdoor and mobile environments in particular present additional difficulties due to camera motion and their variability in backgrounds and lighting conditions. In prior work [7], we presented a mobile VBI that allows control of a wearable computer entirely with hand gesture commands. A collection of recently proposed and novel methods enables hand detection, tracking, and posture recognition for truly interactive interfaces, realized with a head-worn camera and display. For these vision-based hand gesture interfaces it is of tremendous importance to make available a background-invariant, lighting-insensitive, and person- and camera-independent classifier to reliably detect a human's most important manipulative tool, the hand. We will subsequently call these classifiers "detectors."

Hand appearances – the combinations of postures and the directions from which they are viewed – differ in their potential for classification from background and other objects. In order to pick the appearance with the best separability from background (that is, the one that allows detectors to achieve the most robust performance), one could train a detector for each combination and a posteriori analyze their performance. We employ a method that is currently considered the fastest and most accurate pattern detection method for faces in monocular grey-level images (Viola and Jones [10]). Unfortunately, the *training* for this method takes far too long to explore all possible combinations for their suitability to detection.

We introduce an a priori estimate of detector performance. The estimator is based on frequency spectrum analysis and estimates the amount of grey-level variation in the object's appearance. It operates on a prototypical example image of the object to be detected, alleviating the need for extensive data collection. To compare its prediction with actual detection performance, we trained detectors for six hand posture/view combinations. We are interested in the maximal detection rate for a given false positive rate, the "entropy" of the appearance. We found vast differences in detectability with Viola&Jones' method, justifying our methodological approach to find a good VBI initialization gesture. The depth and volume of this study are also novel; we used a total of 2300 hand images. Receiver operating characteristic (ROC) curves are given for all experiments.

The best detector we obtained, combined with skin color verification, achieves outstanding performance in practical application, indoors and outdoors: about one false positive in 100,000 frames. Given that the hand is in the right posture and not extremely over-exposed, it is reliably recognized within a couple frames. This is used to bootstrap the set of subsequent tracking and recognition methods for the wearable interface [7]: the system initializes after the user performs a particular gesture. It then tracks her hand and recognizes a number of key gestures.

The paper is organized as follows. First, we review related work and summarize the Viola-Jones detection method in section 2. Section 3 details our frequency analysis-based estimation of class separability. The actual detector construction, along with details about data collection, evaluation, and performance improvements, is described in section 4. Conclusions are drawn in section 5.

2. Related work

We briefly review approaches to hand and object detection, including the pattern detection method that was the basis for this work.

Most attempts to detect hands from video place restrictions on the environment. For example, skin color is surprisingly uniform [9, 6], so color-based hand detection is possible [13]. However, this by itself is not a reliable modality. Hands have to be distinguished from other skin-colored objects and there are cases of insufficient lighting conditions, such as colored light or grey-level images. Motion flow information is another modality that can fill this gap under certain conditions [2], but for example for non-stationary cameras this approach becomes increasingly difficult and less reliable. Statistical information about hand locations is effective when used as a prior probability [8], but it requires application-specific training. Shape models generally perform well if there is sufficient contrast between the background and the object [1], but they have problems especially with concave objects and cluttered backgrounds. Particle filtering [4] makes shape models more robust to background noise, but shape-based methods are better suited for tracking an object once it has been acquired and they yield only limited results for detection tasks. Cameras that capture depth or thermal infrared images provide additional information that makes hand detection much easier, yet they require specialized and frequently expensive hardware.

Little work has been done on finding hands in grey-level images based on their appearance and texture. Wu and Huang [11] investigated the suitability of a number of classification methods for the purpose of view-independent hand posture recognition. The objective was to classify hand poses, however, so detection performance without the help of skin color information was not considered. Face detection on the other hand has attracted a great amount of interest [12, 3] and many methods relying on shape, texture, and/or temporal information have been described. Texture-based approaches in particular have the potential to yield the best results in varying image environments since they can operate on still images and even cope with partial object occlusions.

2.1. Integration templates

To the best of our knowledge, view-dependent, posture-specific localization of hands in unconstrained grey-level images has not been demonstrated. To achieve this, we use a very fast and accurate learning-based object detection method that was recently proposed and extended by Viola and Jones [10, 5], primarily applied to face detection. It operates on so-called integral images in which each image element contains the sum of the values of all pixels to its upper left, also known as “data cubes” in the database community. This single-pass precomputation step allows for subsequent constant-time summation of arbitrary rectangular areas, or “rectangular features.” During training, “weak” classifiers are selected with AdaBoost, each of them a pixel sum comparison between two or more areas (see fig. 1). Hundreds of these classifiers are then arranged in a multi-stage cascade (termed “detector”), together achieving excellent classification performance. Due to an exhaustive-search component, training a cascade takes on the order of 24 hours on a 30+ node PC cluster.

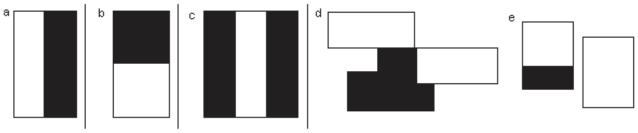

Figure 1: Images a)-c) show one instance of each traditional feature type. For example, type a) can vary in overall width and height as well as in its width ratio of the two rectangular areas, but their heights must not differ. d) and e) are two instances of our new type that allows for almost arbitrary area comparisons since the rectangles’ locations and sizes are less constrained; even overlapping areas are permitted. See also subsection 4.4.

At detection time, the entire image is scanned at multiple scales. For example, a template of size 25x25 pixels, swept across a 640x480 image pixel by pixel, then enlarged in size by 25%, swept again, enlarged, swept, etc. yields 355614 classifications. Every stage of the cascade has to classify the area positive for an overall positive match. This lazy successive cascade evaluation, together with the rectangular features’ constant-time property, allows the detector to run fast enough for the low latency requirements of real-time object detection.

Overall, the method’s accuracy and speed performance, as well as its sole reliance on grey-level images, make it very attractive for hand detection. For practical application outside the context of this paper, we combined it with skin color information for even improved performance.

3. Separability estimation with frequency spectrum analysis

Since training a detector for every possible hand posture (in order to find the best-performing one) is prohibitively expensive, we propose in this section a method to quickly estimate the classification potential, based on only a few training images for each posture. We investigated eight postures

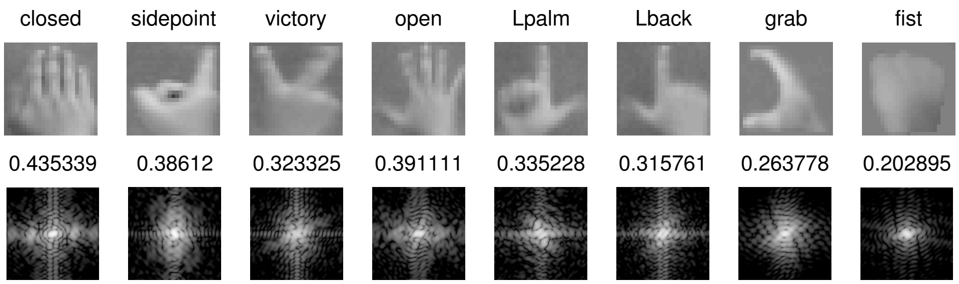

Figure 2: Mean hand appearances and their artifact-free Fourier transforms. Larger s-values (see eq. 4) indicate more high-amplitude frequency components being present, suggesting better suitability for classification from background.

from fixed views, which were selected based on their different appearances and because they can be performed easily. A prototypical example for each posture is shown in fig. 2.

The posture *closed* is a flat palm with all fingers extended and touching each other, *open* is the same but with fingers spread apart. *Sidepoint* is a pointing posture with only the index finger extended, seen from the thumb side. The *victory* or peace posture has index and middle finger extended. The “L” posture involves an abducted thumb and extended index finger and can be seen from the *Lpalm* side and the *Lback* side of the hand. The *grab* gesture is suited to picking up coffee mugs, seen from the top, and the *fist* posture is viewed from the back of the hand.

The separability of two classes depends on many factors, including feature dimensionality and method of classification. There is no known performance estimator for the AdaBoost method described above. Yet it is desirable to a priori predict the potential for successful classification of hand appearances from background due to the detector’s computationally expensive training phase. The estimator presented here is based on the intuition that appearances with a prominent pattern can be detected more reliably than very uniformly shaded appearances. The advantage of the estimator is that it only requires a single prototypical example of the positive class. There is no need for explicit or formal representation of the negative class, “everything else.”

We collected up to ten training images of each of eight hand postures from similar views and computed their mean image (top row in fig. 2). Due to limited training data for *fist* we took only one image and manually set non-skin pixels to a neutral grey. The areas of interest were resized and rescaled to 25x25 pixels, see table 1. The higher-frequency components of a Fourier transform describe the amount of grey-level variation present in an image – exactly what we are looking for. However, the transformation F (eq. 1) introduces strong artificial frequencies, caused by the image’s finite and discrete nature.

$$F(u,v) = \frac{1}{25*25}\sum_{m=0}^{24}\sum_{n=0}^{24} I(m,n)e^{-i2\pi(\frac{mu}{25}+\frac{nv}{25})} \quad (1)$$

We therefore subtract the Fourier transform P of a neutrally colored 25x25-sized image patch from F. This ensures that frequencies resulting from image cropping are eliminated, yielding an artifact-free difference-transform D.

$$D(u,v) = \log|F(u,v) - P(u,v)|, \quad (2)$$

where

$$P(u,v) = \frac{1}{25*25}\sum_{m=0}^{24}\sum_{n=0}^{24}\frac{1}{2}e^{-i2\pi(\frac{mu}{25}+\frac{nv}{25})} \quad (3)$$

In the last step (eq. 4), the sum of all frequency amplitudes is computed, normalized by the Fourier transform’s resolution. This sum is the sought-for estimator, giving an indication of the amount of appearance variation present in the image:

$$s = e^{\frac{1}{k}*\sum_{u,v} D(u,v)}. \quad (4)$$

The bottom row in fig. 2 presents the postures’ artifact-free Fourier transforms D, annotated with s, the sums of their log amplitudes over the entire frequency spectrum. The sums’ absolute values have limited meaning, they are to be regarded in relation to each other. As expected after visual inspection, the *closed* hand appearance has the most amount of grey-level variation, reflected in a high amplitude sum. The *fist*, being mostly a uniformly grey patch, has the least amount of appearance variation, thus also a low s-value.

In the following section, a comparison of the estimates with actual detectors’ performances will confirm our hypothesis – that appearances with larger s-values can be detected more reliably. Computing s-values therefore alleviates the need for the compute-intensive training of many detectors in order to gauge their performance potentials.

4. Detector training and evaluation

This section describes the data collection, training, and evaluation of detectors for the six appearances with the highest s-values. We compare their performances with the posture "detectability" as estimated from their appearance variation, described in the previous section. Furthermore, we optimize training parameters, one improving training speed, the other increasing detection performance. Lastly, the detector chosen for fail-safe VBI-initialization is presented.

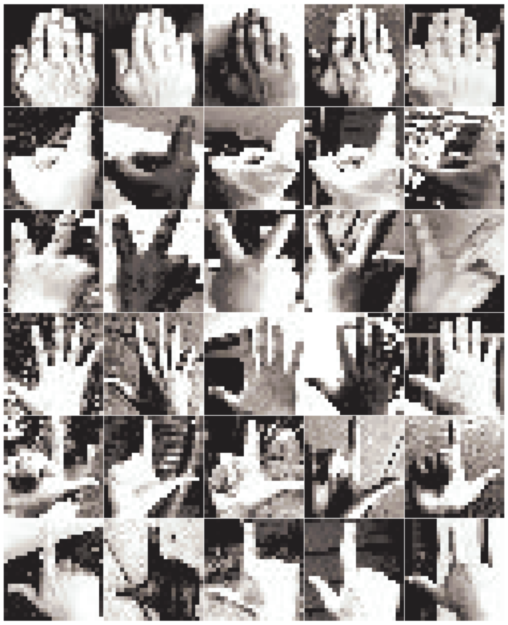

Figure 3: Sample areas of the six hand postures, from top to bottom: *closed, sidepoint, victory, open, Lpalm,* and *Lback*. They are shown in the smallest resolution necessary for detection (25x25).

4.1. Data collection

We collected over 2300 images of hands of ten male and female students' right hands with two different digital still cameras. The pictures were taken indoors and outdoors with widely varying backgrounds and lighting conditions, but without direct sunlight on the hands. The rectangular bounding boxes of the areas containing hand posture appearances were manually marked and rotated to a standard orientation. Figure 3 shows five examples for each of the six postures for which we trained detectors. AdaBoost was performed on one half of the hand images, error rate-validation on the other half (in order to avoid over-training).

The rectangular areas had different but fixed aspect ratios for each of the postures (Table 1). Since we wanted uniform template sizes for all postures for better comparability, this resulted in varying resolutions for the interpolation step. For example, the posture *sidepoint* with a template of size 25 by 25 pixels has twice the sample density along the horizontal dimension than its resolution in the vertical dimension. Similarly, during matching of each detector, different scale factors have to be applied.

closed	sidepoint	victory	open	Lpalm	Lback
389	331	341	455	382	433
0.6785	0.5	0.5	1.0	0.9	0.9

Table 1: The number of training images and the bounding box ratios (width over height) for each posture. Template size and bounding box ratio determine the template resolution along the vertical and horizontal dimensions.

The non-cascaded detectors were trained with more than 23000 negative examples, randomly selected areas from the pictures containing the hand images, but not intersecting the hand areas. Again, half of them were added to the training set, the other half was used for validation. For the cascaded detectors, a pool of 180 random images not containing hands was scanned periodically to dynamically increase the negative training set during training (see ref. [10] for details).

4.2. Non-cascaded detectors

To evaluate predictor accuracy, we first built detectors with unmodified AdaBoost, which produces a single set of weak classifiers for each detector. In sub-section 4.4 we cover cascaded detectors, which are composed of multiple, staged sets of weak classifiers. Here, only the three traditional feature types (see fig. 1) were used.

The detectors were evaluated for their false positive rates by scanning a test set of 200 images not containing hands, some obtained from a web crawl and some taken at our location. Note that the false positive rate is relative to all detector evaluations, and that there are 355614 evaluations required to scan a VGA-sized image (see sub-section 2.1).

Results: The receiver operating characteristic (ROC) curves in fig. 4 show the results of evaluating the six detectors. The posture *closed* fares much better than its competitor hand postures, in that it achieves a higher detection rate for a given false positive rate. This is in line with the prediction of the spectrum-analysis estimator. The *sidepoint* posture does second-best for high detection rates, but then deviates from the prediction. We will see later however that it comparatively does much better again for very low false positive rates with the cascaded detector. Another prediction failure can be observed for the *Lback* and *Lpalm* curves: the more structured *Lpalm* appearance should achieve better class separability. Again, the more expressive features

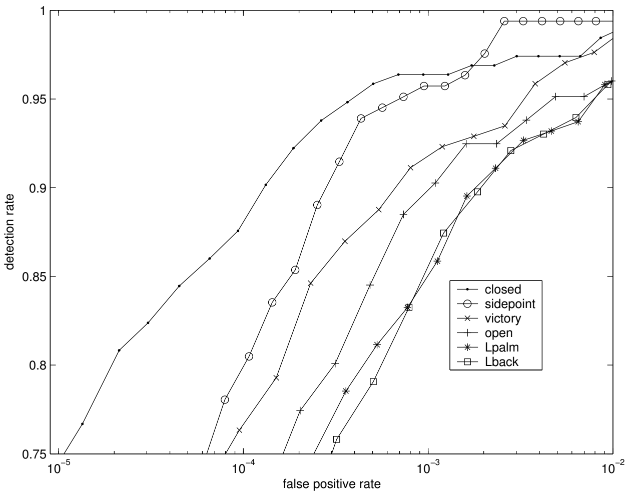

Figure 4: ROC curves for all hand postures, trained on integral images with 25x25 pixel resolution. Each of the six detectors consists of 100 weak classifiers. The x-axis is in log scale.

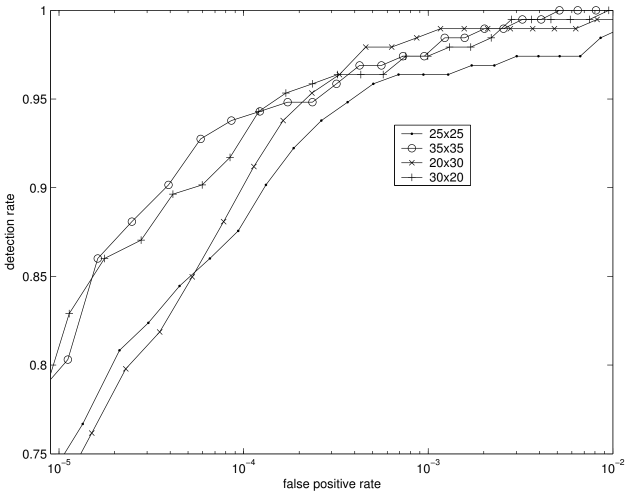

Figure 5: ROC curves for the *closed* posture detector with 100 weak classifiers. The detectors with higher resolution in the horizontal (35x35 and 30x20) outperform the other two.

in the cascaded detector actually do bring out this advantage and are in line with the prediction.

4.3. Effect of template resolution

Before training detectors with more expressive, but also more expensive feature types (on the order of two magnitudes more computational effort during training), we wanted to make sure the integration templates did not contain any redundant information. Therefore, we varied the size of the template area for the best-faring appearance, hoping for resolution reduction without sacrificing accuracy. The impact of different integral image resolutions can be seen in fig. 5.

Unsurprisingly, the finest resolution integral (35x35 pixels) achieves the best performance. Remember that the observed image area is constant, only the sample resolution differs. But higher resolution in the vertical dimension contributes little to this improvement, as witnessed by the lower detection rates of the 20x30 curve. On the other hand, the 30x20 curve has high resolution along the dimension that the estimator frequency analysis showed more high amplitudes for – see the bright horizontal extent in the frequency image for the *closed* posture in fig. 2. This seems to enable the detector to capitalize much more on appearance peculiarities and rewards us with detection rates comparable to the highest-resolution detector.

It is interesting to note that the detector with 30x20 templates performs better for low false positive rates, while a 20x30 resolution performs better for higher false positive rates. We speculate that the stretch in the vertical produces large, uniform areas that allow for easy distinction between hands and *many* other appearances. The lack in horizontal resolution however compresses away the fine finger structures that are required for separation from *most* other appearances.

4.4. The final detector

In this sub-section we present the final result of our research, a hand detector with a very low false positive rate. We also show that the particular choice of feature types influences the relative detectability of hand appearances.

For each posture, we trained a cascaded detector that could select its weak classifiers from a set of four feature types instead of from only three types as were used in ref. [10] and in section 4.2. The novel feature type is a comparison of four rectangular areas. During training, they can move about relative to each other with “no strings attached,” even partially overlapping each other, just their sizes are restricted (see fig. 1). These more powerful features allow the detector to achieve better accuracy, demonstrated in fig. 6.

Results: The relative performance of detector pairs stays roughly the same, even though the curves are not as smooth as with non-cascaded detectors due to the staged cascading and the resulting evaluation method (details in Viola and Jones’ paper [10]). Of particular interest are the left parts of the curves since a fail-safe hand detection for vision-based interfaces must be on the conservative side. There, the cascaded detectors show ROCs along the lines of the performance predicted in section 3: *closed* outperforms all others, *sidepoint* is second-best, and the more structured appear-

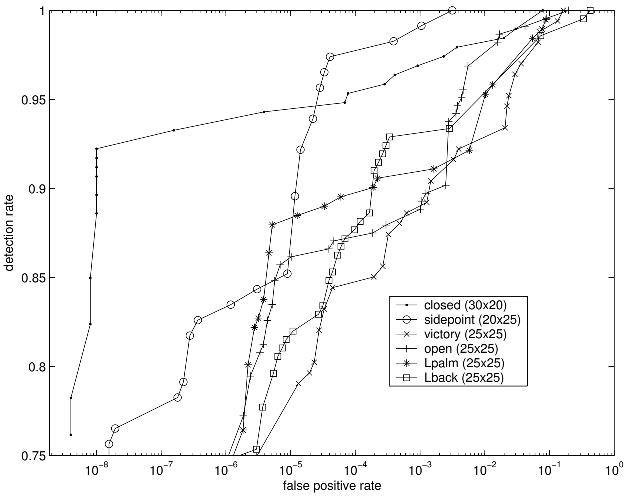

Figure 6: ROC curves for cascaded detectors with the less constrained feature type. Note that the scale on the y axis is different from previous figures.

ance *Lpalm* now does better than the more uniform *Lback*.

Extrapolating from the results of this study, we suggest that mostly convex appearances with internal grey-level variation are better suited to the purpose of detection with the rectangle feature-classification method. The *open* posture for example already has a lower Fourier structure value, hinting that background noise hinders extraction of consistent patterns. The detector's accuracy confirms the difficulty to distinguish hands from other appearances.

The final hand detector that we chose for our application detects the *closed* posture. For scenarios where we desire fast detection, we picked the parameterization that achieved a detection rate of 92.23% with a false positive rate of $1.01 * 10^{-8}$ in the test set, or one false hit in 279 VGA-sized frames. For most scenarios it is sufficient however to pick a parameterization that had a detection rate of 65.80%, but not one false positive in the test set. The high frame rate of the algorithm almost guarantees that the posture is detected within a few consecutive frames.

5. Summary and conclusions

Computer Vision methods for hand gesture interfaces must surpass current performance in terms of robustness and speed to achieve interactivity and usability. Recent advances in pattern recognition have made detection at frame rate possible. We investigated the suitability of various hand postures for fail-safe detection before arbitrary backgrounds in grey-level images.

Our contributions are as follows. First, we demonstrate the suitability of the integral-image approach to the task of detecting hand appearances. Second, a qualitative measure is presented that amounts to an a priori estimate of "detectability," alleviating the need for compute-intensive training. Third, parameters of the detection method are optimized, achieving significant speed and accuracy improvements. Overall, this study shows how the Viola-Jones detector can achieve excellent detection rates for hand postures. These results provide an important step towards easily deployable and robust vision-based hand gesture interfaces.

References

[1] T. F. Cootes and C. J. Taylor. Active Shape Models: Smart Snakes. In *Proceedings of the British Machine Vision Conference*, pages 9–18. Springer-Verlag, 1992.

[2] R. Cutler and M. Turk. View-based Interpretation of Real-time Optical Flow for Gesture Recognition. In *Proc. IEEE Intl. Conference on Automatic Face and Gesture Recognition*, pages 416–421, April 1998.

[3] E. Hjelmås and B. K. Low. Face Detection: A Survey. *Computer Vision and Image Understanding*, 83(3):236–274, September 2001.

[4] M. Isard and A. Blake. Condensation – Conditional Density Propagation for Visual Tracking. *Int. Journal of Computer Vision*, 1998.

[5] M. Jones and P. Viola. Fast Multi-view Face Detection. Technical Report TR2003-96, MERL, July 2003.

[6] M. J. Jones and J. M. Rehg. Statistical Color Models with Application to Skin Detection. *Int. Journal of Computer Vision*, 46(1):81–96, Jan 2002.

[7] M. Kölsch, M. Turk, T. Höllerer, and J. Chainey. Vision-based Interfaces for Mobility. Technical Report TR 2004-04, University of California at Santa Barbara, February 2004.

[8] T. Kurata, T. Okuma, M. Kourogi, and K. Sakaue. The Hand Mouse: GMM Hand-color Classification and Mean Shift Tracking. In *Second Intl. Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems*, July 2001.

[9] D. Saxe and R. Foulds. Toward robust skin identification in video images. In *Proc. IEEE Intl. Conference on Automatic Face and Gesture Recognition*, pages 379–384, Sept. 1996.

[10] P. Viola and M. Jones. Robust Real-time Object Detection. *Int. Journal of Computer Vision*, 2002.

[11] Y. Wu and T. S. Huang. View-independent Recognition of Hand Postures. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 84–94, 2000.

[12] M.-H. Yang, D. J. Kriegman, and N. Ahuja. Detecting Faces in Images: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 1 2002.

[13] X. Zhu, J. Yang, and A. Waibel. Segmenting Hands of Arbitrary Color. In *Proc. IEEE Intl. Conference on Automatic Face and Gesture Recognition*, 2000.