Analysis of Rotational Robustness of Hand Detection with a Viola-Jones Detector

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

Abstract

The research described in this paper analyzes the in-plane rotational robustness of the Viola-Jones object detection method when used for hand appearance detection. We determine the rotational bounds for training and detection for achieving undiminished performance without an increase in classifier complexity. The result – up to 15° total – differs from the method's performance on faces (30° total). We found that randomly rotating the training data within these bounds allows for detection rates about one order of magnitude better than those trained on strictly aligned data. The implications of the results effect both savings in training costs as well as increased naturalness and comfort of vision-based hand gesture interfaces.

1. Introduction

In previous work [5], we investigated various hand postures and views for their suitability to reliable detection with arbitrary backgrounds. Using Viola and Jones' extremely fast pattern recognition method (Section 3 and [8]), we found vast differences in the achievable accuracy, that is, the maximum detection rate for a given false positive rate. We used a detector for the best suited posture in a vision-based hand gesture interface, where robust hand detection is of utmost importance. This vision interface serves as the sole input modality to control all functions of a wearable computer system [6]. The detector, combined with skin color verification, has *not once* detected a false positive in practical application, indoors and outdoors, and in many hours of operation. Given that the hand is in the right posture, it is recognized within a few frames. The high detection accuracy allows reliable initialization of a set of dependent computer vision methods that track the hand and recognize key postures, despite the instability of a head-worn camera and the unknowns of uncontrolled environments.

Unfortunately, said object detection method is not inherently invariant to in-plane object rotations, requiring the user of our mobile system to perform very precise gestures – a daunting task with a head-worn camera. Viola and Jones extended their method to detect objects exhibiting arbitrary in-plane rotations, requiring additional effort both algorithmically, during training, and during detection [3].

Our objective for this work was to analyze the limits of the original approach [8], without incurring a performance penalty, and for objects other than faces (because their appearance characteristics are entirely different, see Section 3.1). We explain the method, introduce a new feature type, and discuss the dataset in Section 3. Section 4 presents our experiments and the results. We show that detection accuracy can be improved by an order of magnitude without algorithmic modifications, while the speed performance remains unchanged. These results are consistent for a number of hand postures and appearances. We employed the improved detectors in our mobile vision interface and can report better and faster initialization due to more natural and less rigid hand postures required for detection.

2. Related work

Face detection has attracted a great amount of interest [10, 2] and many methods relying on shape, texture, and/or temporal information have been described. Texture-based approaches in particular have yielded the most universal results. However, little work has been done on finding hands in unconstrained grey-level images, not even view-dependent, posture-specific hand appearances. Instead, most attempts to detect hands from video place restrictions on the environment. For example, skin color is surprisingly uniform [7, 11, 4], so color-based hand detection is possible. This by itself is not reliable as hands have to be distinguished from other skin-colored objects. Motion flow information is another modality that can fill this gap [1], but for non-stationary cameras this approach becomes quite complex. An extensive study [9] investigated the suitability of a number of classification methods for the purpose of view-independent hand posture recognition. However, detection without the help of skin color information and real-time performance were not considered.

Our work investigates unimodal hand detection in unconstrained grey-level images with a method that meets the real-time requirements of vision-based interfaces (VBI).

3. Viola-Jones detector

The basis for this work is a learning-based object detection method, recently proposed by Viola and Jones [8]. It

is considered the fastest and most accurate pattern recognition method for faces in monocular grey-level images, and in prior work we confirmed similarly excellent performance for hand detection [5]. The method operates on so-called *integral images*: each image element contains the sum of all pixels values to its upper left, allowing for constant-time summation of arbitrary rectangular areas. During training, "weak" classifiers are selected with AdaBoost, each of them a pixel sum comparison between rectangular areas. Since the originally proposed feature types did not have sufficient discriminative power for all hand appearances, we added a new feature that achieved superior classification rates. This feature (Fig. 1) is more general than the one introduced in [3]. Hundreds of these classifiers are then arranged in a multi-stage cascade. Lazy successive cascade evaluation and the constant-time property allow the detector to run fast enough for the low latency requirements of VBIs. The method's accuracy, speed performance, and its sole reliance on grey-level images make it very attractive for hand detection.

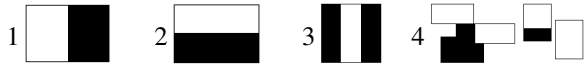

Figure 1: The feature types. The fourth type (shown in two instances) compares two white areas with two black areas. Its areas' locations and sizes are only minimally constrained and can even overlap.

3.1. In-Plane Rotations

Viola-Jones face detectors handle about 30° of in-plane rotation of frontal and profile views, 15° in either direction [3]. However, we found detectors for hands to be much more sensitive to in-plane rotations (see Subsection 4.1); this prompted the research presented here. The difference to faces stems from the hand's smaller features (fingers) being more sensitive to correct alignment during training, as well as from less inter-person appearance variation of a certain posture and view.

Jones and Viola recently extended their method to detect arbitrary in-plane rotations and side views of faces [3]. In a first stage of classification, implemented with a decision tree, one of twelve classifiers is selected. Each of these handles detection of faces within about 30° of in-plane rotations. While this approach is still very fast, it adds training time and about doubles detection time.

Similarly, we investigated detection of in-plane rotations of various hand postures. However, our focus was not on covering the entire 360° range of rotations but instead to increase each detector's range of detected rotations without adding any computational overhead and without negatively affecting the false positive rate.

3.2. Dataset, Training, and Baseline

We created a training set of over 330 very well aligned hand appearances for each of six posture/view combinations from different people's right hands, taken with different cameras in both indoor and outdoor settings (Fig. 2). One classifier was trained for every posture on half the images and validated on the other half. 180 random images not containing hands were scanned to periodically increase the negative training set (see [8] for details).

Figure 2: The six postures, 25x25 pixels, bottom row rotated by 15°. From left: *closed*, *open*, *sidepoint*, *victory*, *Lpalm*, and *Lback*.

For the *closed* posture, we rotated both the training and validation sets by various amounts around the image area's center and trained one classifier for each angle. Consistent parameters for the training caused equally-complex cascade stages throughout all experiments in this paper. The evaluation (Fig. 3) shows that there are no large differences in the accuracy of the classifiers, especially for low false positive rates. Establishing this baseline is important because some rotations could be intrinsically harder to detect than others – these experiments dismiss this possibility.

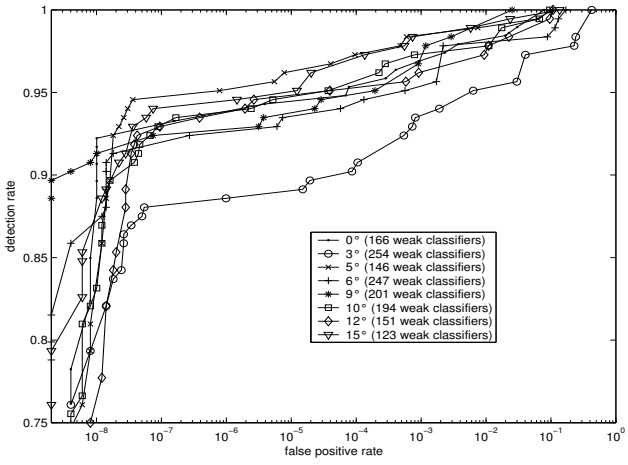

Figure 3: Classifiers trained and evaluated for the same rotation.

4. Rotational robustness

4.1. Problem: Rotational Sensitivity

To demonstrate the sensitivity of the detection method when used for hand appearances, we tested a classifier trained on

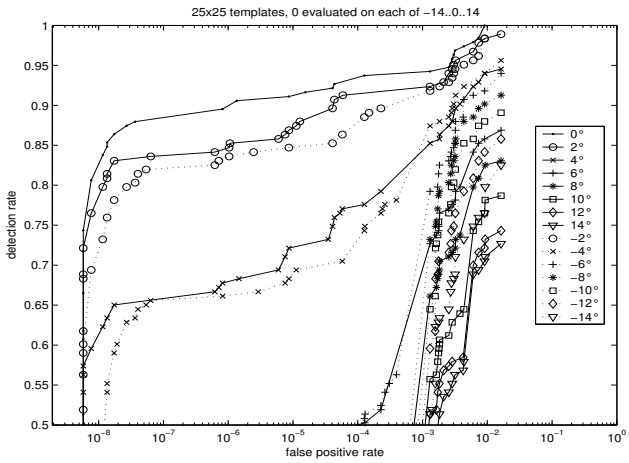

Figure 4: Trained for unrotated training images, evaluated for test images rotated by various angles. There is a sharp decrease in detection accuracy for in-plane rotations of 4° or more. Note the symmetry for rotations to the left and right.

well-aligned examples for its accuracy. In contrast to Viola and Jones' face detector, we found poor accuracy with rotated test images for as little as 4° (Fig. 4). A second set of experiments shows that this is not caused by peculiarities of the unrotated appearance of the particular hand posture: Eight classifiers were trained on examples rotated by various degrees, then tested with examples rotated randomly between 0° and 15°. The results in Fig. 5 demonstrate their high rotational sensitivity in contrast to a classifier that was trained on 0°-15°-rotated examples (top curve; the difference is even larger for false positive rates below 10^{-4}).

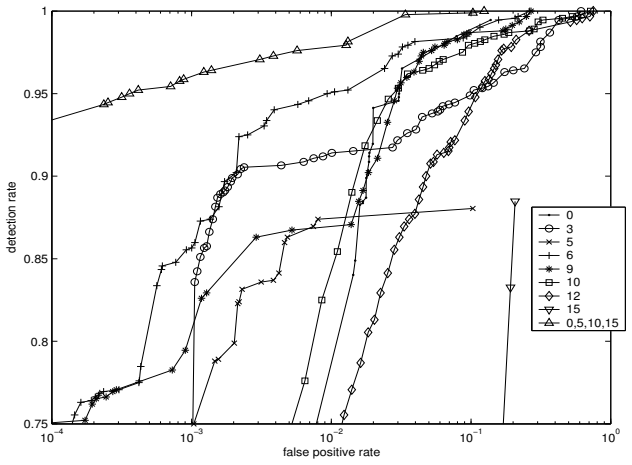

Figure 5: Trained for stated angle, evaluated on randomly rotated test set. None of the fixed-angle classifiers achieves accuracy close to the classifier trained for various angles.

4.2. Bounds

The objective of the second set of experiments was to determine the angles we could rotate the training examples and still achieve good detection performance on the equally-rotated test set. We created a large training set with four repetitions of the same images of the *closed* posture, each rotated by an additional 15°. A Viola-Jones detector over time keeps the positive examples that are reliably detectable, while it successively ignores those that would require an unacceptably high false positive rate. The experiment's assumption is that well-detectable examples will be retained and all others sacrificed in order to achieve a low false positive rate. The evaluation (Fig. 6) suggests that the examples with 0° and 15° of rotation are more consistently recognizable than those with 30° and 45°. Therefore, we set the bounds for rotating the training examples to within 15°.

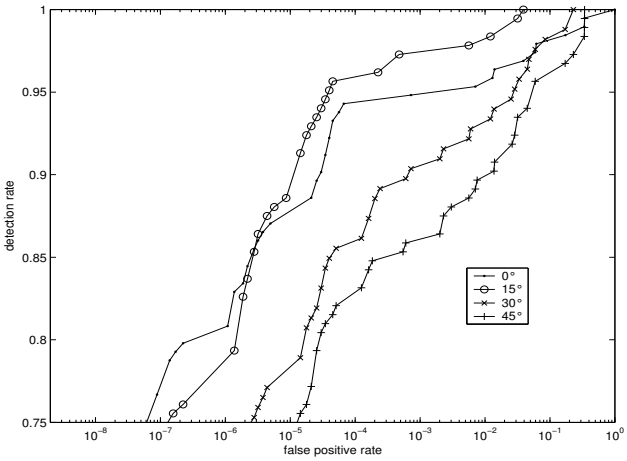

Figure 6: A classifier created on a training set with multiple rotations does not treat all angles equally. Instead, examples rotated by 30° and 45° are more likely to be dropped in favor of examples with smaller rotations.

4.3. Rotation Density of Training Data

Next, we were interested in the influence that different rotation angle densities have on training and detection performance. Three classifiers were trained; their training and validation sets contained examples rotated in varying steps: A = {0, 5, 10, 15}, B = {0, 3, 6, 9, 12, 15}, and C = {0..15} with random angles. They consisted of 198, 190, and 239 weak classifiers, respectively. The detectors were evaluated on examples randomly rotated between 0 and 15 degrees.

No significant accuracy variation was observed. We conclude that accuracy is not affected by rotation angle density for angles of 5° or less. This is an important result because wider steps allow for fewer training examples, reducing data collection effort and computational training cost.

4.4. Other Postures

Finally, we confirmed the applicability of the main results obtained for the *closed* posture to the other five postures. Fig. 7 plots the detection rate of classifiers built with rotated training sets (0°-15° random) divided by that of classifiers built with unrotated training sets. Both were evaluated with a test set with all examples rotated by 15°. The detectors trained on rotated examples achieve at least equal performance, and for low false positive rates they outperform the detectors trained on fixed examples by about one order of magnitude. They also have a lower minimum false positive rate while still detecting some hand appearances.

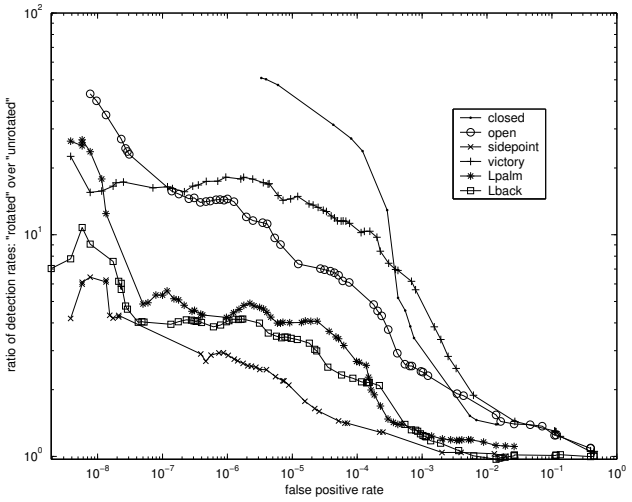

Figure 7: Ratio of detection rates for "trained on rotated" over "trained on unrotated", evaluated on 15° rotated areas. There are no data points if the unrotated classifiers have a detection rate of zero.

4.5. Overall Results and Discussion

The number of weak classifiers required for a certain accuracy did not differ significantly between classifiers trained on rotated and unrotated training images. Since we used consistent training parameters (number of weak classifiers per cascade stage and their accuracy) for all detectors, the resulting detection speed performance of classifiers for 0°-15°-rotated images is about equal to that of classifiers that detect unrotated images only.

The results presented in this paper are likely to generalize to other objects because the surveyed hand appearances exhibit very different characteristics, such as their convexity (*open* vs. *closed*), their texture variation (*Lback* vs. *Lpalm*), and the background to foreground ratio (*closed* vs. *sidepoint*). As detailed in Section 4.1, presenting training images that are rotated within these bounds is crucial to good accuracy for object appearances other than faces.

5. Conclusions

This paper's contribution is a detailed analysis of Viola-Jones detectors for in-plane rotations of hand appearances. The main result is that only about 15° of rotations can be efficiently detected with one detector, different from the method's performance on faces. Most importantly, the training data must contain rotated example images within these rotation limits. Detection rates on rotated appearances improve by about one order of magnitude, without a negative impact on detection speed. We also introduce a more expressive feature type that is required for good accuracy on some hand appearances.

6. Acknowledgments

This work was partially supported under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under contract No. W-7405-ENG-48.

References

[1] 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.

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

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

[4] 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.

[5] M. Kölsch and M. Turk. Robust Hand Detection. In *Proc. IEEE Intl. Conference on Automatic Face and Gesture Recognition*, May 2004.

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

[7] 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.

[8] P. Viola and M. Jones. Robust Real-time Object Detection. In *Intl. Workshop on Statistical and Computational Theories of Vision*, July 2001.

[9] 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.

[10] 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.

[11] B. D. Zarit, B. J. Super, and F. K. H. Quek. Comparison of Five Color Models in Skin Pixel Classification. In *Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, pages 58–63, Sept. 1999.