ABSTRACT
For gesture and sign language recognition, hand shape and hand motion are the primary sources of information that differentiate one sign from another. Building an efficient and reliable hand detector is therefore an important step in recognizing signs and gestures. In this paper we evaluate three hand detection methods on three sign language data sets: a skin and motion detector [1], hand detection using multiple proposals [12], and chains model [9].

Categories and Subject Descriptors
H.5.2 [User Interfaces]: Input Devices and Strategies; I.4.8 [Scene Analysis]: Object Recognition

General Terms
Experimentation

Keywords
Hand detection

1. INTRODUCTION
In the computer vision community, hand detection has been a subject of study for several years, due to its obvious applicability in domains such as sign language recognition, gesture recognition, and human-computer interfaces. Accurate detection of hands in still images or video is still a challenging problem, due to the variability of hand appearance. Since hands do not have a fixed shape, their shape is difficult to describe computationally. This is in contrast to faces, for example, which have a well-defined shape (with two eyes, a nose, a mouth), and thus can be easily detected with commercial products such as cameras and cell phones. Colored gloves and magnetic trackers can give accurate detection results, but they are expensive and inconvenient, since users have to wear special equipment.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

PETRA ’14, May 27 - 30 2014, Island of Rhodes, Greece
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2746-6/14/05...$15.00
http://dx.doi.org/10.1145/2674396.2674442

In this paper we evaluate three hand detection methods on three sign language data sets: a skin and motion detector [1], hand detection using multiple proposals [12], and chains model [9].

2. RELATED WORK
Methods for detecting hands using computer vision can be categorized into four groups: (1) appearance-based hand detection, (2) detecting hands as part of human pictorial structure, (3) hand tracking and (4) hand shape detection.

2.1 Appearance-Based Hand Detection
Some methods use skin color information to localize and track hands in signing video [7, 6]. Mittal et al. [12] describe a method for detecting hands and their orientation using skin color, hand shape, and context. Kölsch et al. [10] study view-specific hand posture detection with an object recognition method proposed by Viola and Jones [18]. Ong et al. [14] present a novel, unsupervised approach to train an efficient and robust detector which is capable of not only detecting the presence of human hands in an image but also of classifying the hand shape. Zhong et al. [20] evaluate four features for hand detection: color, temporal motion, gradient norm, and motion residue.

2.2 Detecting Hands as Part of Human Pictorial Structure

2.3 2D Hand Tracking
Yuan et al. [19] propose a temporal filtering framework for hand tracking. In each frame, simple features like color and motion residue are exploited to identify multiple candidate hand locations. The temporal filter then uses Viterbi algorithm to select among the candidates from frame to frame.
Trinh et al. [17] use binary quadratic programming to integrate appearance, motion and complex interaction between the hands. Morariu [13] describes a framework that uses probabilistic and deterministic networks and their AND/OR search space to detect and track the hands and feet of multiple interacting persons from a single camera view.

2.4 Hand Shape Detection
Athitsos et al. [2] propose a method for detecting shapes of variable structure in images with clutter. Variable structure means that some shape parts can be repeated an arbitrary number of times, some parts can be optional, and some parts can have several alternative appearances. A new class of shape models is introduced, called Hidden State Shape Models, that can naturally represent shapes of variable structure. A detection algorithm is described that finds instances of such shapes in images with large amounts of clutter by finding globally optimal correspondences between image features and shape models. Thayananthan et al. [16] compare two methods for object localization from contours: shape context and chamfer matching of templates.

3. SKIN AND MOTION DETECTOR
3.1 Detecting Skin
Since human skin is relatively uniform in color, a statistical color model can be employed to compute the probability of every pixel being skin color. In [8], a skin color likelihood distribution and a non-skin color distribution, denoted as $P(r, g, b|\text{skin})$ and $P(r, g, b|\neg\text{skin})$, respectively, are proposed, in which the RGB color space is quantized to $32 \times 32 \times 32$ values. Based on these two distributions, the probability of a pixel, whose color vector is $[r g b]$, being skin is defined using Bayes rule:

$$P(\text{skin}|r, g, b) = \frac{P(r, g, b|\text{skin})P(\text{skin})}{P(r, g, b)}$$  \hspace{1cm} (1)

3.2 Temporal Motion
Motion information is another useful cue for hand detection in gesture videos, since a user needs to move at least one hand to perform a hand gesture.

To detect motion, we used a simple method based on frame differencing. Other more sophisticated background subtraction methods such as Mixtures of Gaussian (MoGs) can be used instead, but the simple frame differencing method has worked sufficiently well in our experiments.

Frame differencing works as follows: let $I(x, y, i)$ denote the intensity value at pixel $(x, y)$, at the $i$-th frame. By comparing $I(x, y, i)$ with $I(x, y, i - z)$ and $I(x, y, i + z)$, we compute a motion indicator value $M(x, y, i)$. Motion indicator $M(x, y, i)$ is defined using the following equations:

$$I_1(x, y, i) = |I(x, y, i) - I(x, y, i - z)| \hspace{1cm} (2)$$

$$I_2(x, y, i) = |I(x, y, i) - I(x, y, i + z)| \hspace{1cm} (3)$$

$$M(x, y, i) = \min(I_1(x, y, i), I_2(x, y, i)) \hspace{1cm} (4)$$

For every pixel in a frame, we can compute the skin indicator value and motion indicator value of the pixel using the methods described above. Let $S(x, y, i)$ denote the skin indicator value at pixel $(x, y)$, in the $i$-th frame and $M(x, y, i)$ denote the motion indicator value at pixel $(x, y)$, in the $i$-th frame. The combined skin and motion indicator value for this pixel is defined using the following equation:

$$A(x, y, i) = S(x, y, i) \times M(x, y, i) \hspace{1cm} (5)$$

The most likely hand candidate is defined as the region which has the largest summation of values in image $A$. Once we find a candidate region, before we identify the next most likely region, we overwrite with value zero the skin color probabilities of all pixels in the candidate region we have just identified. This helps to avoid identifying multiple candidate regions with significant mutual overlap. To determine the hand region’s size, at every frame, the size of the hand is adjusted according to the size of the face. Faces are detected using the Viola-Jones method [18].

![Video frames](image)

Figure 1: Video frames of two example signs from DS2 and DS3.

4. EXPERIMENTS
We conducted our experiments in a user independent manner using three sign language video datasets—DS1, DS2 and DS3. Information about the number of videos and the number of frames, separated into one-handed and two-handed instances can be found in table 1. Figure 1 shows frames from two example sign language videos.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>DS1</th>
<th>DS2</th>
<th>DS3</th>
</tr>
</thead>
<tbody>
<tr>
<td># of one-handed video</td>
<td>42</td>
<td>42</td>
<td>42</td>
</tr>
<tr>
<td># of one-handed frame</td>
<td>902</td>
<td>1276</td>
<td>1197</td>
</tr>
<tr>
<td># of two-handed video</td>
<td>48</td>
<td>48</td>
<td>48</td>
</tr>
<tr>
<td># of two-handed frame</td>
<td>1337</td>
<td>1945</td>
<td>1735</td>
</tr>
</tbody>
</table>

For our experiments, we evaluated three methods. The first was the multiple proposals method of [12], which uses a combination of scores from a skin detector, context detector, and
Figure 2: Detection accuracy on DS1, DS2 and DS3 data sets from top to bottom. The left side is the result on one-handed signs and the right side on two-handed signs. The x-axis corresponds to the top $k$ (from 2 to 20) hand candidates, while y-axis corresponds to detection accuracy.
hand shape detector to produce hand bounding box proposals. This method uses an external training set, so all DS1, DS2, and DS3 videos are used for testing.

The second method was the chains method of [9], which generates chains connecting a known object—the face, for example—to the object of interest (i.e. the hands). Each dataset is tested separately, using one of the others as a training set. DS2 was used to train the system for DS1 experiments, DS3 was used to train for DS2 experiments, and DS1 was used for experiments with DS3.

The final method we evaluated was the skin and motion-based detection of [1]. It does not require a training set, so all videos were utilized for testing.

4.1 Measure of Accuracy

The hand detection is considered to be correct if it is within a half-face width from the ground-truth location of the hand. We report the detection performance within the top $k$ ($k$ is from 1 to 20 for one-handed case and from 2 to 20 for two-handed case) hand candidates per frame (Figure 2). So if the ground truth is within half the face width of one of the top $k$ candidates, it is considered accurately located. The one-handed and two-handed test cases are examined separately.

Figure 2 is the evaluation result. The left side is the result on one-handed videos and the right side is the result on two-handed videos. And the first row, second row and third row corresponds to result on DS1, DS2 and DS3 respectively.

5. DISCUSSION

It can be seen from figure 2 that the skin and motion detector consistently outperforms the chains model and multiple proposals methods on both one-handed and two-handed signs. As an example, for one-handed signs in the DS3 data set, the top candidate from the skin and motion detector method provides an accurate hand location for over 80% of the frames, whereas the other methods achieve less than 30% accuracy. Performance of the skin and motion methods drops to just over 70% for the top 2 candidates, however, on the two-handed signs in the same data set.

5. DISCUSSION

5. DISCUSSION

5. DISCUSSION

6. CONCLUSIONS AND FUTURE WORK

This paper has compared three hand detection methods [1, 12, 9] on three sign language data sets. By comparing these three popular hand detection methods, we can see that the skin and motion based method provides the best results on our sign language data sets. It is also clear, however, that its performance on two-handed signs drops considerably. In future work, we will explore more sophisticated features, such as HOG [5] and motion residue [19], and try a more advanced algorithm, such as adaboost, to combine these features. And at the same time, we will utilize tracking algorithms to ensure hand candidate temporal consistency across frames, rather than relying on single frame detection.

7. ACKNOWLEDGMENTS

This work was partially supported by National Science Foundation grants IIS-1055062, CNS-1059235, CNS-1035913, and CNS-1338118.

8. REFERENCES


