Experiments with Computer Vision Methods for Hand Detection

Zhong Zhang, Rommel Alonzo, and Vassilis Athitsos
Computer Science and Engineering Department
University of Texas at Arlington, USA

ABSTRACT

For gesture and sign language recognition, hand shape and hand motion are the primary sources of information that differentiate one sign from another. So, building an efficient and reliable hand detector is an important step for recognizing signs and gesture. In this paper we evaluate four features for hand detection: color, temporal motion, gradient norm, and motion residue, and we explore the potential of these features for building a reliable hand detector. At first, we use these four features separately to identify where the hands are in each frame of our gesture videos. Then we evaluate different combinations of such features using weighted linear combination, so to build a more accurate hand detector. Experimental results show the relative performance of the four features in isolation and in different combinations, and demonstrate promising results for detectors that combine these features.

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, feature combination

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. Hands do not have a fixed shape, and thus their shape is hard 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 detected these days by commercial products such as cameras and cell phones. Colored gloves and magnetic trackers can give accurate detection results, but they are expensive and inconvenient(users have to wear special equipment).

In this paper, we evaluate four features (color, temporal motion, gradient norm and motion residue). Using those features we design seven hand detectors. First, four features are used separately to identify the bounding box of hands in each frame of our sign videos. A statistical skin color model is employed to compute the probability of every pixel being skin-like color. Hand candidates are identified as rectangle areas with the highest sum of probabilities. Temporal motion is calculated by comparing the current frame with the previous and posterior frame in gray scale. Gradients are computed using Gaussian smoothing followed by a simple 1-D $[-1, 0, 1]$ mask. For motion residue, for every two consecutive frames, the first frame is partitioned into blocks and then the best match of each block in next frame is found by translation. Based on the best match of each block, motion residue is defined as the average of L2 distance in intensity level between the block and its best match in the next frame. We tested these four features in seven ways: in addition to the four features used separately, we evaluate a combination of skin and motion, a combination of skin, motion and gradient norms, and a combination of skin, motion and motion residue.

2. RELATED WORK

Several approaches have been proposed for hand detection in recent years. There are two major groups. One of them relies on the apperance of the hand itself. The other utilizes the context supplied by surrounding parts. Karlinsky et al. [8] developed a ’chains model’ which can locate parts of interest in a robust and precise manner. In their model, the relation between context features and the target part is modeled in a non-parametric manner using an ensemble of feature chains leading from parts in the context to the detection target. Buehler et al. [5] detect hands and arms at the same time in sign language videos. They cast the problem as inference in a generative model of the image.

The appearance based methods exploit features from only the hands, without using information from other body parts. Zhu et al. [14] generates a hand color model and a background color model for a given image, and then uses these models to classify each pixel in the image as either a hand pixel or a background pixel. Ong et al. [10] presents a novel, unsupervised approach to training an efficient and robust detector which not only detects the presence of human hands within an image but also classifies the hand shape. In their paper, a tree structure of boosted cascades is constructed. The head of the tree provides a general hand detector while the individual branches of the tree classify a valid shape as belong to

PETRA ’11, May 25-27, 2011, Crete, Greece.
Copyright 2011 ACM ISBN 978-1-4503-0772-7/11/05 ...$10.00.

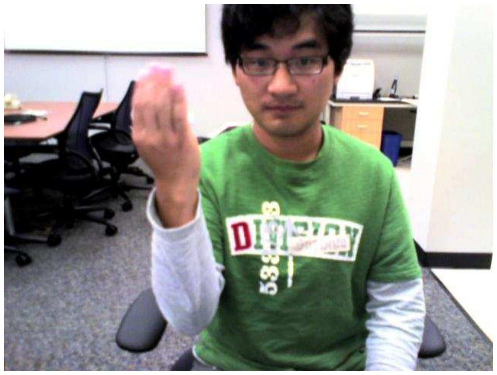

Figure 1: An example frame from a video of a gesture.

one of the predetermined clusters exemplified by an indicative hand shape. Kölsch et al. [9] presented a view-specific hand posture detection with an object recognition method recently proposed by Viola and Jones. Yuan et al. [13] propose using motion residue as a feature for hand detection, and use dynamic programming to identify an optimal sequence of hand locations in a video.

Another category of methods represents the shape of hands as a deformable shape, and uses shape matching algorithms to find hands. Athitsos et al. [2] proposed Hidden State Shape Models to represent shapes of variable structure. Hand shapes are modeled as shapes of variable structure, and 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. [11] compare two methods for object localization from contours: shape context [4] and chamfer matching of templates [3]. Coughlan et al. [6] presented a novel deformable template which detects the boundary of an open hand in a grayscale image without initialization by the user.

3. SEVEN VISION-BASED METHODS FOR HAND DETECTION

3.1 Detecting hands based on color information

Since the human skin is relatively uniform, a statistical color model can be employed to compute the probability of every pixel being skin color. In [7], a skin color likelihood distribtion and a non-skin color distribution, denoted as $P(r,g,b|skin)$ and $P(r,g,b|\neg skin)$ respectively are proposed, in which the color space is in RGB but quantized to 32*32*32 values. Based on these two distributions, the probability of a pixel, whose color vector is $[rgb]$, being skin is defined using Bayes rule:

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

Figure 2 shows an image visualizing the $P(skin|r,g,b)$ score computed for every pixel of the video frame shown on Figure 1.

The most likely hand region is defined as the region which has the largest summation of posterior skin color probability over the candidate region. 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 us avoid identifying multiple candidate regions with significant mutual overlap. As for how 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 [12].

3.2 Detecting hands based on temporal motion

Motion information is another discriminant 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 have used a simple method based on frame differencing. More sophisticated background subtraction methods, e.g., Mixtures of Gaussian(MoGs), can be used instead, but the simple frame differencing method has worked sufficiently well for our experiemtns so far.

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 value $M(x,y,i)$ is defined using the following equations:

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

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

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

Figure 2 shows an image $M(x,y,i)$ computed for the video frame shown on Figure 1.

The most likely hand candidate is defined as the region which has the largest summation of values in the M image. As we did with the skin feature, once we find a candidate region, we set zero to all items of this candidate region, to suppress producing multiple candidate regions with too much overlap.

3.3 Detecting hands based on gradients

Before computing the gradients, we use a Gaussian mask to smooth every frame. And then the gradients are computed using a simple 1-D $[-1,0,1]$ mask. Let $I(x,y,i)$ denote the intensity value at pixel (x,y), at the i-th frame. The gradient at this pixel, $G(x,y,i)$ is computed using the following equations:

$$dx = I(x-1,y,i) - I(x+1,y,i) \tag{5}$$

$$dy = I(x,y-1,i) - I(x,y+1,i) \tag{6}$$

$$G(x,y,i) = \sqrt[2]{dx^2 + dy^2} \tag{7}$$

Figure 2 shows an image $G(x,y,i)$ computed for the video frame shown on Figure 1.

The most likely hand candidate is defined as the region which has the largest sum of gradient norm values over the candidate region, similar to the approach we followed for the skin and motion features.

3.4 Detecting hands based on motion residue

Hands typically undergo non-rigid motion, because they are deformable articulated objects. This means that hand apperance changes more frequently from frame to frame, compared to apperance of other background objects. We can use this property to detect hands,

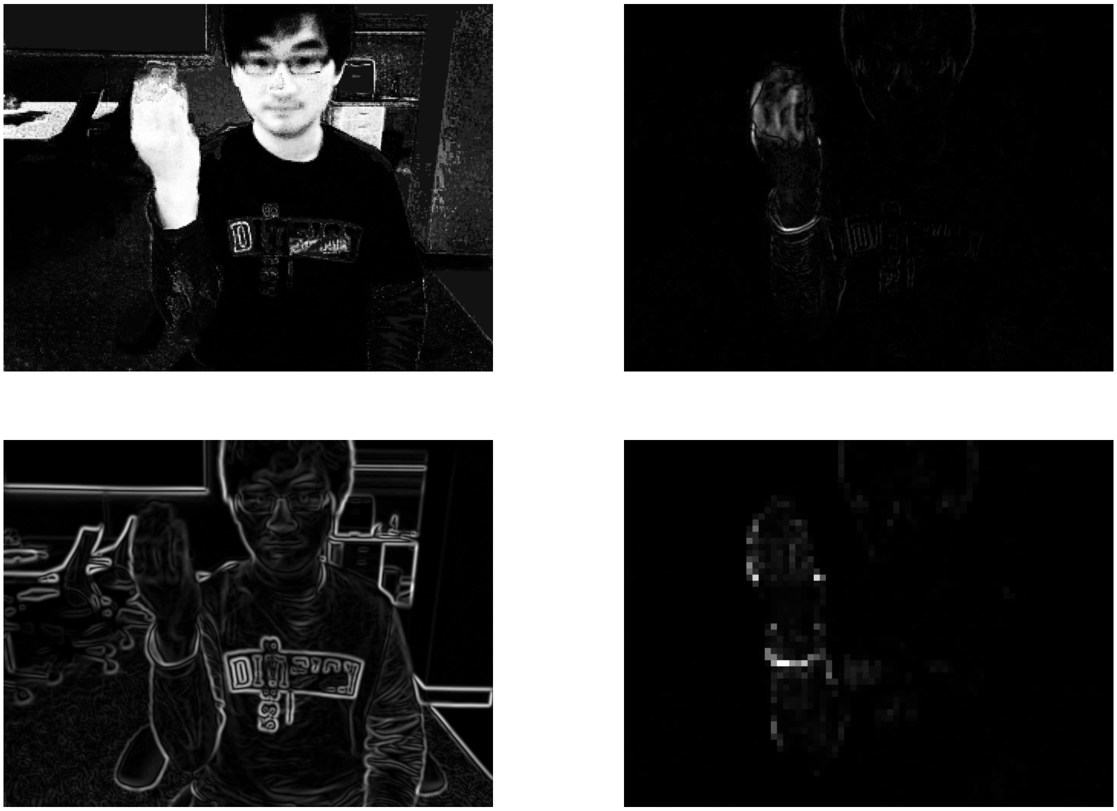

Figure 2: Scores computed based on skin (top left), motion (top right), gradient norms (bottom left), and motion residue (bottom right) features, for the original image shown on Figure 1.

by identifying regions in each frame that have no good matches(in terms of apperance) among regions in the next frame. This is an idea proposed in [13].

Following the approach of [13], for every two consecutive frames, the first frame is partitioned into blocks(8x8) and then we try to find best match of each block in the next frame by translation. Based on the best match of each block, motion residue is defined as the summation of differences in intensity level between the block and its best match in the next frame. Let A is the one block in current frame and B is the best match of A in the next frame. We can use the following equation to calculate residue:

$$R = \sum_{i \in \text{block}} (A_i - B_i)^2 \tag{8}$$

Every pixel in the block will be assigned as this residue. Because hands move nonrigidly in most cases, the blocks in a hand region tend to have high residues, and therefore we can use residue as a feature to detect hands. Hand candidates are identified as rectangle areas with the largest residue value.

Figure 2 shows an image of motion residue scores computed for the video frame shown on Figure 1.

3.5 Detecting hands based on feature combinations

3.5.1 Skin and temporal motion

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

$$A(x, y, i) = S(x, y, i) * M(x, y, i) \tag{9}$$

3.5.2 Skin, temporal motion and gradients

For every pixel in one frame, we can compute the skin indicator value, motion indicator value and gradient indicator value using the methods introduced before. Let $S(x, y, i)$ denote the skin indicator value at pixel (x, y), at the i-th frame, $M(x, y, i)$ denote the motion indicator value at pixel (x, y), at the i-th frame and $G(x, y, i)$ denote the gradient indicator value at pixel (x, y), at the i-th frame. We combine these three features using the following equation:

$$B(x, y, i) = e^{\ln(S(x,y,i)+\epsilon)+\ln(M(x,y,i)+\epsilon)+w_1*\ln(G(x,y,i)+\epsilon)} \tag{10}$$

ϵ is a small positive constant and the weight w_1, which is $\frac{1}{20}$ is chosen by searching over many possible values, so as to optimize performance on the training data.

3.5.3 Skin, temporal motion and motion residue

For every pixel in one frame, we can compute the skin indicator value, motion indicator value and motion residue indicator value using the methods introduced before. Let $S(x, y, i)$ denote the skin indicator value at pixel (x, y), at the i-th frame, $M(x, y, i)$ denote the motion indicator value at pixel (x, y), at the i-th frame and $R(x, y, i)$ denote the motion residue indicator value at pixel (x, y), at the i-th frame. We combine these three features using the following equation:

$$B(x, y, i) = e^{\ln(S(x,y,i)+\epsilon)+\ln(M(x,y,i)+\epsilon)+w_2*\ln(R(x,y,i)+\epsilon)} \tag{11}$$

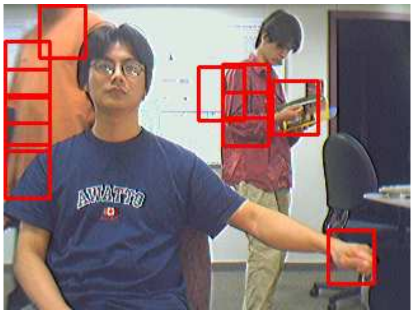

Figure 3: Top ten candidates of skin, motion and residue method for one frame

ϵ is a small positive constant and the weight w_2, which is $\frac{1}{2}$is chosen by searching over many possible values, so as to optimize performance on the training data.

4. EXPERIMENTS

There are three experimental datasets, the “easy digits” data set, the “hard digits” data set, and the “ASL” data set. The “easy digits” dataset includes 240 videos, 14885 frames in total, 14885 hands in total. The “hard digits” dataset includes 140 videos, 7996 frames in total, 7996 hands in total. The ASL dataset includes 91 videos, 3065 frames in total. We use the easy digits data set as our training set, and the other two datasets as test sets. The “easy digit” and “hard digit” datasets are publicly available [1].

4.1 Measure of Accuracy

For every frame, we use every method introduced before to find the top ten candidates. These ten candidates are the most possible bouding boxes which include hands. Actually, we also annotate our data set manually so that we have the exact bouding box for every hand in every frame. The L2 distance between centroid of hand rectangle and centroid of every candidate bounding box is computed and the minimum one is chosen as the result of current frame. If the current frame has two hands, we have two minimum results. The last step is to normalize the results based on the diagonal of face, where face is detected using Adaboost method [12]. In this way, for hard digits data set, we have 7996 values and for American Sign Language data set, we have 3065 values. We set some percentiles of hands, saying $[0.6 : 0.05 : 0.9, 0.91 : 0.01 : 0.99]$ in our experiment. And every percentile corresponds to one distance. The distance is defined as the minimum value that make the percentile of hands’ value less than this minimum value.

Figure 5 and Figure 6 are the evaluation results. Figure 3 shows the top ten candidates for one frame using skin, motion and residue mothod. Figure 4 shows the top ten candidates for one frame using skin and motion method.

5. DISCUSSION

This paper has described ongoing work towards a system for hand detection. We explore the potential of four features for hand detection, as well as the potential of combinations of those four features. In the future, we will try more sophisticated features, such as HOG and LBP to see what accuracy we can get. And we will also

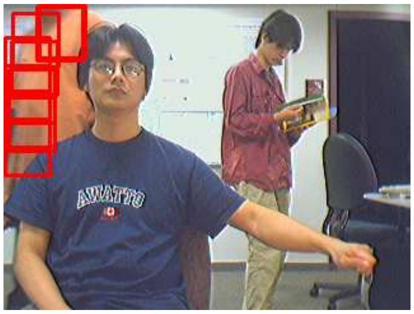

Figure 4: Top ten candidates of skin and motion for one frame

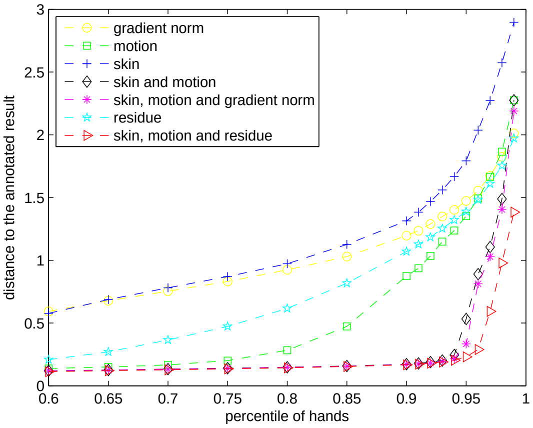

Figure 5: Detection result plot for hard digits data set. The x-axis corresponds to the percentiles of hands, while the y-axis corresponds to distance to the annotated resutl.

employ a more advanced algorithm to combine these features, such as Adaboost.

Acknowledgements

The research reported here has been partially funded by grants from the National Science Foundation: IIS-0705749, IIS-0812601, CNS-0923494, and CNS-1035913. This work has also been supported by a UTA startup grant to Professor Athitsos, and UTA STARS awards to Professors Chris Ding and Fillia Makedon.

6. REFERENCES

[1] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff. A unified framework for gesture recognition and spatiotemporal gesture segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 31(9):1685–1699, 2009.

[2] V. Athitsos, J. Wang, S. Sclaroff, and M. Betke. Detecting instances of shape classes that exhibit variable structure. In *European Conference on Computer Vision (ECCV)*, pages 121–134, 2006.

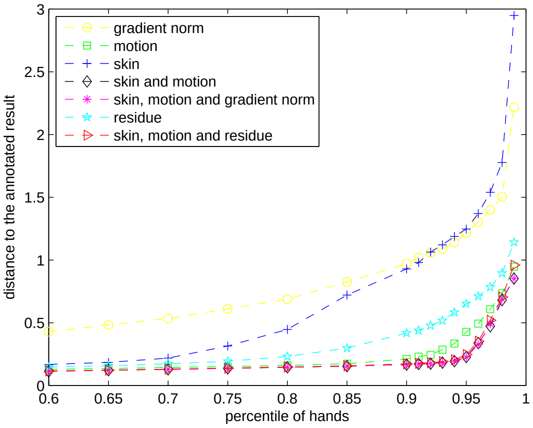

Figure 6: Detection result plot for uta data set. The x-axis corresponds to the percentiles of hands, while the y-axis corresponds to distance to the annotated resutl.

[3] H. Barrow, J. Tenenbaum, R. Bolles, and H. Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *International Joint Conference on Artificial Intelligence*, pages 659–663, 1977.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 24(4):509–522, 2002.

[5] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman. Long term arm and hand tracking for continuous sign language TV broadcasts. In *British Machine Vision Conference(BMVC)*, 2008.

[6] J. Coughlan, A. Yuille, C. English, and D. Snow. Efficient deformable template detection and localization without user initialization. *Journal of Computer Vision and Image Understanding*, 78, 2000.

[7] M. Jones and J. Rehg. Statistical color models with application to skin detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages I:274–280, 1999.

[8] L. Karlinsky, M. Dinerstein, D. Harari, and S. Ullman. The chains model for detecting parts by their context. In *IEEE International Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 25–32, 2010.

[9] M. Kölsch and M. Turk. Robust hand detection. In *IEEE International Conference on Automatic Face and Gesture Recognition(AFGR)*, pages 614–619, 2004.

[10] E. J. Ong and R. Bowden. A boosted classifier tree for hand shape detection. In *Face and Gesture Recognition*, pages 889–894, 2004.

[11] A. Thayananthan, B. Stenger, P. H. S. Torr, and R. Cipolla. Shape context and chamfer matching in cluttered scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 127–133, 2003.

[12] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 511–518, 2001.

[13] Q. Yuan, S. Sclaroff, and V. Athitsos. Automatic 2D hand

tracking in video sequences. In *IEEE Workshop on Applications of Computer Vision*, pages 250–256, 2005.

[14] X. Zhu, J. Yang, and A. Waibel. Segmenting hands of arbitrary color. In *IEEE International Conference on Automatic Face and Gesture Recognition(AFGR)*, pages 446–455, 2000.