# Multi-Scale Topological Features for Hand Posture Representation and Analysis

Kaoning Hu and Lijun Yin
State University of New York at Binghamton
khu1@binghamton.edu, lijun@cs.binghamton.edu

## Abstract

*In this paper, we propose a multi-scale topological feature representation for automatic analysis of hand posture. Such topological features have the advantage of being posture-dependent while being preserved under certain variations of illumination, rotation, personal dependency, etc. Our method studies the topology of the holes between the hand region and its convex hull. Inspired by the principle of Persistent Homology, which is the theory of computational topology for topological feature analysis over multiple scales, we construct the multi-scale Betti Numbers matrix (MSBNM) for the topological feature representation. In our experiments, we used 12 different hand postures and compared our features with three popular features (HOG, MCT, and Shape Context) on different data sets. In addition to hand postures, we also extend the feature representations to arm postures. The results demonstrate the feasibility and reliability of the proposed method.*

## 1. Introduction

Hand gesture is one of the most commonly used signals for human communication, expression, and demonstration [14]. The complex structure of hands and the large number of hand variations in various scales, occlusions, and rotations make it a very challenging task for automatic analysis.

In general, hand gestures express certain information by either dynamic movement [22, 24] (i.e., temporal hand gestures) or static shape or configuration (i.e., static hand postures) [14], and can involve one hand or two hands [16]. In this paper, we focus on the static hand postures of a single hand.

Various approaches have been previously applied for the feature representations and recognition of hand postures. These features can be based upon shape, e.g. Shape Context [2] and Chamfer Distance [26]; texture, e.g. Modified Census Transform (MCT) [10]; or both shape and texture, e.g. Histogram of Oriented Gradient (HOG) [3, 17].

In shape-based approaches, hand postures are recognized by matching the hand contour to the training samples or to a model built upon the training samples. The 2D hand models [13] are B-spline curves constructed by control points which are matched to the hand region by using partitioned sampling algorithms. However, a single 2D model is limited to describe one or two postures [13]. The 3D models [5, 18] are articulated to model the knuckles, finger tips and the wrist. A Kalman Filter [18] is used to match the hand image to the 3D models. However, Kalman Filter is restricted to cases with known backgrounds [15]. Some methods rely on the robust tracking of feature points such as knuckles, and may require manual correction to these points [5]. Oikonomidis successfully used Particle Swarm Optimization to track 2 interacting hands, but the tracking frame rate is 4Hz, and a set of RGB-D camera is required [16]. A common advantage of the shape based approaches is that many of these features are invariant to rotation and scaling [3]. In contrast, the approaches with texture included are more robust when the background is cluttered [3]. However, texture features are not invariant to rotations, which restrains the range of hand movement. It is still very challenging to robustly describe and recognize the variety and individuality of hand features. Due to the complexity of hand postures and the desire of real-time applications, people either fuse multiple features [25] or develop new features. In this paper, we develop a novel multi-scale topological feature representation inspired by Persistent Homology.

Persistent Homology is a method for analysis of homology at different scales [27]. It is an algebraic topological invariant that has been used as a mathematical tool for algorithmically analyzing topological features of shapes or functions, and has been previously applied to the problem of shape analysis and retrieval [4, 7]. However, those shapes are quite more distinguishable than various hand postures. Moreover, the topological similarity of hand postures makes it difficult to distinguish various hand postures only based on the topology of hand shapes.

In this work, we consider the complementary holes between the hand region and its convex hull as topological spaces. We show that by examining these spaces, highly discerning topological features are obtained. This finding leads us to explore a new feature representation of hand

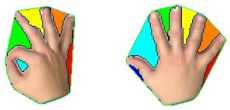

Figure 1. The complementary holes (colored regions) between the hand postures and their convex hulls.

postures, which lies in the hand's complementary holes. Specifically, we track the existence of the holes' topological features as the scale changes. We propose a multi-scale topological feature representation for hand posture analysis inspired by the principle of Persistent Homology. We study the holes between the hand region and its convex hull and distinguish different hand postures by the unique topology of these holes. The unique multi-scale Betti Numbers matrix (MSBNM) is computed for hand posture representation, characterization, and classification.

## 2. Topological Feature Representation of Hand Posture

### 2.1. Representing hand postures using complementary holes

Topological features rely on the number of parts (connected components) and holes of an object. Those features are distinguishable for different objects while preserved to distortions of the same object.

Since many hand postures yield the same topology as they do not show any holes, we propose to "create" topological holes from hand images through the construction of their convex hulls. The *convex hull* of a given hand posture is the minimal convex set containing the hand region and can be derived by a convex polygon. Due to the typical concave shape of the hand region, non-trivial holes are observed between the hand region and its convex hull. We call these holes the *complementary holes* of a hand posture. Figure 1 shows two examples of different postures and their complementary holes. Using the complementary holes to describe hand postures has the following advantages. First, it is observed that each hand posture produces unique sets of complementary holes. Second, the complementary holes are independent to the illuminations and rotations. Third, it is straightforward to quantify each set of the complementary holes using the topological features.

Although each hand posture produces unique sets of complementary holes, simply counting the number of holes is not sufficient to distinguish different postures, nor reliable enough to tolerate variances of a single posture. This is because the number of holes of a posture is changeable due to noise in the hand images or variations of viewing angles. Ambiguity may be caused by the same number of holes from different postures. Multi-scale analysis for shape description could be a remedy to alleviate this problem [4]. Inspired by the Persistent Homology, we propose a multi-scale topological analysis to address this issue.

### 2.2. Persistent Homology and Betti Numbers

Homology is a mathematical tool used to algorithmically analyze topological features of shapes. In other words, it is an algebraic procedure for counting "parts", "holes", and "voids" of various types [8]. Betti numbers [27] is used to quantify the topological features: In a topological space, B0 is defined by the number of parts (connected components) of the space, and B1 is defined by the number of holes.

Persistent Homology is the analysis of homology at different scales. A homology group K can be connected by a series of homomorphism $(H_i, i = 1, 2, ..s)$ processes with scale 1 to $s$:

$$H_1(K) \rightarrow H_2(K) \rightarrow ... \rightarrow H_s(K) \rightarrow 0 \qquad (1)$$

In Persistent Homology, the scale is associated with the connectivity of the points. Two points are considered as connected if the distance between them is smaller than a threshold [6, 27], and the scale is defined by the threshold. So this scaling process will change the connectivity of the points, therefore the persistence of the holes can change, and so can the topological features. Because topological features come in all scale-levels and can be nested or in more complicated relationships, observing the homology classes as to how they change as the scale changes [6] will help us to exploit detailed features of hand postures.

To serve our purpose of hand posture representation, the number of holes with respect to sequential scales is adopted for Betti number representation. If we group the hand shape and its convex hull as one topological space, we can analyze the number of holes (B1) of this space. Moreover, if we consider each hole as a topological space, we can analyze the number of parts (B0) of the hole's space (i.e. a hole may split to multiple holes) as the scale changes. Such a representation is not only taking account of the number of holes but also the life-span of each hole. Details about the feature representation of hand postures are illustrated in the following sections.

### 2.3. Feature representation by MSBNM

Persistent Homology can detect holes in a coordinate-free system [27]. In order to instantiate its application on a Cartesian coordinate system, which is the 2D domain on which the hand images are represented, we simplify the computation of Persistent Homology with image processing tools to locate holes in 2D images.

In addition, the definition of scaling in Persistent Homology allows us to take advantage of the computational simplicity of the morphological operations to "connect" different components and represent different scales. In other

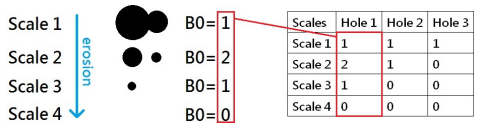


Figure 2. Construction of the B0 sequence, and B0 sequence being part of multi-scale Betti Numbers matrix

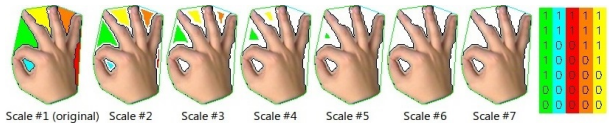


Figure 3. Construction of multi-scale Betti Numbers matrix

words, dilating the image is a way of implementing the scaling process used in Persistent Homology computations. Multiple morphological dilations of the hand region leads to multiple scales associated with the homology (i.e., 2D holes) groups of the hand image. Since we consider each hole as a topological space, applying the dilation operation to the convex hull region of a hand is equivalent to applying an erosion operation $O$ to the hole regions of the hand within the convex hull. This operation is applied repeatedly until the holes are completely eroded.

$$O_1(K) \rightarrow O_2(K) \rightarrow ... \rightarrow O_s(K) \rightarrow 0 \qquad (2)$$

In our proposed approach, we track the topology of each hole as the scale changes. At the original scale, each hole is shown as an individual region of connected component and is considered as a topological space. As the scale changes, the hole can stay, split, or vanish. The existence (e.g., life-span) of each hole is described using a B0 sequence. If the hole stays at one scale, its B0 value is 1. If it splits, its B0 value is increased accordingly. If it vanishes, its B0 value is 0. Figure 2 illustrates the construction of B0 sequence as the scale changes.

A B0 sequence is constructed for each hole in the convex hull of the hand. We use the multi-scale Betti Numbers matrix (MSBNM) as the feature representation to classify different hand postures. The multi-scale Betti Numbers are defined as a matrix as shown in Figure 2, where each column represents the B0 sequence (connected components) of each hole along with the scales at which it is found, as we consider each hole as a topological space. Each row is associated with a certain scale. Figure 3 illustrates the construction of the multi-scale Betti Number matrix. Notice that the sum of each row is the total number of holes at each individual scale, which is also the hand's B1 value of the topological space within its convex hull.

By computing multi-scale Betti Numbers (MSBNM), we are capable to neutralize the effect of noise. If a hole is short-lived, it is likely just caused by noise. However, if a

Figure 4. Examples of multi-scale Betti Numbers matrices of different hand postures. Each column of the matrix is the B0 sequence of the corresponding hole highlighted by the same color.

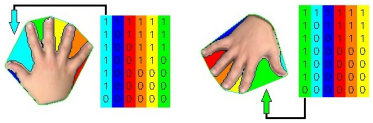

Figure 5. Hole correspondence.

hole is preserved at very large scales, the hole is significant and should be considered as a key feature of the hand posture. Moreover, if a hole can split into multiple holes, it must have a special shape (for example, like a dumbbell).

As a result, MSBNM partially reflects the size of each hole, improving the distinguishability of different postures with a same topology.

Figure 4 shows MSBNM of three hand postures from the dorsal view. The holes are highlighted with different colors. Ideally, the scale should change continuously, but we only compute a fixed number of scales to reduce computation redundancy. After intense experiments, we chose 7 scales, including the original scale. In short, this new feature representation is used to construct a unique feature matrix to classify certain hand postures. Since we only choose 7 instead of infinite scales, some large holes could stay in all scales.

## 2.4. Acquisition of MSBNM

In this subsection, we address the issues how to construct scale and rotation invariant multi-scale Betti Numbers matrix (MSBNM), and how to maintain the consistent dimension of MSBNM, followed by the steps of the acquisition of MSBNM.

As previously mentioned, morphological erosion is employed to represent different scales. In order to build a set of scale-invariant Betti Numbers, the size of the erosion structures must also be adaptive to the size of the hand region. We bound the hand region using an ellipse $E$, and use a smaller ellipse which is 1:32 (by dimension, not area) of $E$ as the erosion structure. This parameter yields the best performance in our experiments.

Since a hand can rotate to arbitrary angle, we need to establish a correspondence to construct rotation invariant MSBNM. The hole with longest life-span is determined as the reference hole. This hole is the most significant and reliable hole. Then, the rest of the holes are sorted in counter-

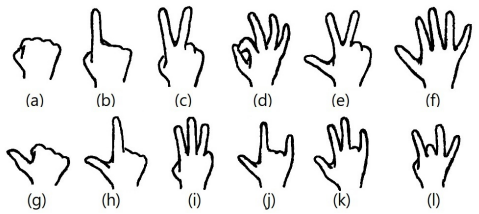


Figure 6. Twelve different hand postures

clockwise order with respect to the center of the hand, following the reference hole. Figure 5 shows an example where the holes of two samples with the same posture were matched.

Based on our observation, hand postures can only generate a finite number of hole regions. Thus, for hand posture analysis, we only count at most 7 largest holes, resulting in 7 columns of MSBNM. In cases where some postures produce less than 7 holes, we add padding columns of 0s to the MSBNM to make the matrix dimensions consistent.

The acquisition of MSBNM follows six steps:

(1) Compute the convex hull of the hand region.

(2) Consider the hand region and its convex hull as one topological space and locate all holes inside the convex hull. It is optional to use a threshold and discard the tiny holes which are caused by noise.

(3) Enclose the hand region using an ellipse $E$, and use a smaller ellipse which is $\frac{1}{32}$ of $E$ as the erosion structure.

(4) Consider each hole as a topological space, and apply the morphological erosion operations iteratively 6 times on the region to generate multiple scale representations. Then, we record the B0 (connected components) sequence of each individual hole space along with each step of the erosion operation.

(5) Take the longest lived hole as the reference hole, and put its B0 sequence into the first column of MSBNM.

(6) Sort the remaining holes in counter-clockwise order with the reference hole first. Put their B0 numbers into the corresponding columns of MSBNM. Padding columns of 0s are added if the matrix has less than 7 columns.

Then, the MSBNM is to be used as the input of a classifier for hand posture classification.

## 3. Experiments on Hand Posture Recognition

We created a database of 12 different types of hand postures with 100 samples of each posture and 1,200 samples in total. The 12 postures are shown in Figure 6. In addition, we have also tested our approach on another public hand posture database [20].

Before the hand posture is described by MSBNM, the region of the hand must be detected and segmented. There exist different approaches to track and extract hand regions

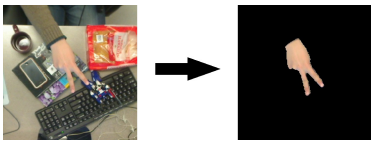

Figure 7. The segmentation of hand region

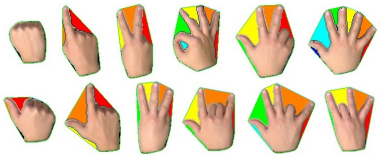

Figure 8. Samples of our dataset with convex hull detected.

from images [11, 21]. Some researchers have also developed reliable hand tracking system with the assistance of color gloves [23]. Since our focus is to evaluate the efficacy of the new hand feature representation, we use a simple yet relatively reliable method [9], which combines background subtraction, skin region segmentation, and the AAM method to detect the hand region. Further to ensure a stable background, a regular color camera is placed above the hand. Figure 7 shows the segmentation of hand region.

After segmentation of the hand region, the program computes the MSBNM of the hand posture at each frame. In this experiment, all samples were collected from the dorsal view under the same illumination. The yaw rotation range of the hand is $[0°, 90°]$, and the pitching/rolling of the hand is $[-30°, 30°]$. The resolution of the image is $720 \times 480$, while a hand region is approximately $150 \times 150$. Twelve samples of our dataset are shown in Figure 8.

The performance of our MSBNM approach has been compared to existing feature types, including HOG [17], MCT [10], and Shape Context [2]. Note that an alternative approach based on the Chamfer Distance [19] is not included for comparison as it has been proven rotation/scale dependent [19] and is very time consuming [1] for practical applications.

For MSBNM, MCT, and HOG, we train each of them on 3 different classifiers, Decision Tree (DT), Bayesian Network (BN), and Classification via Regression (CR). The *performance* of each feature is represented by the optimum accuracy across all classifiers. For example, when using MSBNM, Decision Tree yields the highest accuracy, so we use this accuracy to represent the *performance* of MSBNM. The *performance* of Shape Context is evaluated using the Nearest-Neighbor based classifier suggested in its references [2]. In our experiments, we used 5-fold cross-validation on our data set (1200 samples). The accuracies of each feature representation are shown in Table 1.

The confusion matrix of the classification by Decision Tree using MSBNM is shown in Table 2.

| Approaches | DT | BN | CR | *Performance* |
|---|---|---|---|---|
| MSBNM (ours) | **94.8** | 85.7 | 94.3 | **94.8** |
| MCT [10] | 83.4 | **88.6** | **88.6** | **88.6** |
| HOG [17] | 86.6 | 82.6 | **90.1** | **90.1** |
| Shape Context [2] | N/A | N/A | N/A | **85.0** |

Table 1. Accuracy comparison of each feature representation, evaluated using cross-validation on our dataset. Shape Context is evaluated by the classifier in [2].

| a | b | c | d | e | f | g | h | i | j | k | l | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **100** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | a |
| 0 | **98** | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | b |
| 0 | 0 | **93** | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 6 | c |
| 0 | 0 | 0 | **93** | 0 | 0 | 0 | 0 | 3 | 4 | 0 | 0 | d |
| 0 | 0 | 0 | 0 | **98** | 0 | 0 | 0 | 0 | 0 | 1 | 1 | e |
| 0 | 1 | 0 | 0 | 0 | **96** | 0 | 0 | 1 | 0 | 2 | 0 | f |
| 0 | 0 | 0 | 0 | 0 | 0 | **100** | 0 | 0 | 0 | 0 | 0 | g |
| 0 | 3 | 2 | 0 | 0 | 0 | 0 | **81** | 0 | 0 | 0 | 14 | h |
| 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | **97** | 0 | 0 | 0 | i |
| 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | **92** | 2 | 3 | j |
| 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 1 | **93** | 3 | k |
| 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | **97** | l |

Table 2. The confusion matrix (shown as percentages) of 12 classes, evaluated using cross-validation on our dataset. Each row represents the samples of the same class.

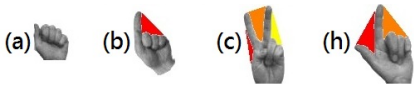


Figure 9. Samples of Jochen's dataset.

From Table 1, we can see our MSBNM approach outperforms HOG, MCT, and Shape Context.

Besides our own data set, we also tested on a dataset of Jochen Triesch Static Hand Posture Database [20]. Since the focus of this work is not on hand region segmentation, we only used the data set with light background. We selected 4 types of postures from their data set, which had been included in our posture set as (a), (b), (c) and (h) shown in Figure 6. As a result, 96 samples, collected from 24 different persons, were used for testing. The samples of their postures are shown in Figure 9.

Since the hand images of Jochen's dataset are taken from palm view, where the textures are different from dorsal view, we created a training set of 100 samples of each of the 12 hand postures (1200 in total) of the palm view from our lab. The accuracies of classification of the Jochen's dataset are shown in Table 3.

The confusion matrix of the classification by Decision Tree using multi-scale Betti Numbers is shown in Table 4.

From above tables, we can see that our MSBNM approach outperformed all of the other methods even when we used our database for training and the Jochen's dataset for testing. It is observed that topological features are better

| Approaches | DT | BN | CR | *Performance* |
|---|---|---|---|---|
| MSBNM (ours) | **84.4** | 72.9 | **84.4** | **84.4** |
| MCT [10] | **25.0** | **25.0** | **25.0** | **25.0** |
| HOG [17] | **25.0** | **25.0** | **25.0** | **25.0** |
| Shape Context [2] | N/A | N/A | N/A | **33.3** |

Table 3. Accuracy comparison of each feature representation, trained on our dataset and tested on Jochen's dataset.

| a | b | c | d | e | f | g | h | i | j | k | l | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **83** | 13 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | a |
| 0 | **92** | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | b |
| 0 | 0 | **63** | 0 | 0 | 0 | 0 | 37 | 0 | 0 | 0 | 0 | c |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | **100** | 0 | 0 | 0 | 0 | h |

Table 4. The confusion matrix (shown as percentages) of 4 classes, trained on our dataset and tested on Jochen's dataset. Each row represents the samples of the same class.

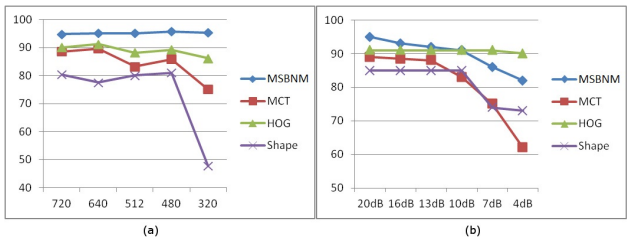


Figure 10. Performance curves with respect to different resolutions (a) and different SNRs (b)

conserved than the texture and hand contour. The results demonstrate that our approach is more robust than the compared approaches in terms of the personal-independent test for hand posture classification.

## 4. Performance Evaluation

In order to evaluate the robustness of our MSBNM approach, we conducted experiments for hand posture recognition under various imaging conditions (e.g., different image qualities, rotations, illuminations, etc.), and compared it to the state of the art approaches (e.g., MCT, HOG, and Shape Context).

### 4.1. Evaluation on various image resolutions

We evaluated the performance using hand images with different resolutions. We conducted experiments on various image resolutions, i.e., 720p, 640p, 512p, 480p, and 320p, respectively. The *performance* (highest accuracy across different classifiers) at each resolution is plotted in Figure 10 (a).

As shown in Figure 10 (a), the performance of MSBNM is relatively stable and superior to the other approaches from the high resolution to the low resolution. It shows that the proposed topological feature representation is more robust under various image resolutions than the compared ap-

proaches.

### 4.2. Evaluation on degraded images with noise

We evaluated the performance using hand images with noise. There are two types of degradations affecting the hand postures recognition. The first one is the Gaussian noise caused by the hardware. The second one is imperfect or noisy segmentation of the hand region caused by uneven illumination. In our experiments, we evaluated the performance by adding random image degradations of both types to our data set used in Section 3. We then compared our approach to the other approaches. We controlled the SNR of each image quality level, and plotted the *performance* (highest accuracy across different classifiers) curve of each approach regarding to the SNR, shown in Figure 10 (b).

In Figure 10 (b), the performance of our approach is superior to MCT and Shape Context. It is also superior to HOG unless the noise is increased dramatically (e.g. SNR becomes lower than 10dB). MCT computes texture features of a hand image, so it is deteriorated by the Gaussian noise. The degradation of our approach and Shape Context is caused by imperfect segmentation of the fingers due to the presence of heavy noise.

### 4.3. Evaluation under cross illuminations

Knowing that illumination conditions can impact the performance of hand posture recognition, we evaluated the proposed MSBNM approach using a training set and a testing set with different lighting conditions. We collected hand posture samples under four different lighting conditions: bright illumination, medium bright illumination, dark illumination, and uneven illumination. Besides changes to the pixel intensities, inconsistent illuminations may also cause poor white balance and Gaussian noise from the camera. Thus hand region segmentation may be poor and contain errors due to the lower quality images and result in a degradation in recognition performance.

In our experiments, under each illumination condition, we collected 100 samples of each posture. The data set consisted of 4,800 samples in total. Then, we performed a 4-fold cross-validation, where each fold contained samples of one specific illumination condition. Note that the data with three lighting conditions were used for training, and the remaining data with a fourth lighting condition was used for testing. Therefore the illumination condition in the test set was different with those used in the training set. We repeat different combinations using 4-fold cross-validation. This experiment is referred to as *cross illuminations* and the accuracies are shown in Table 5.

As shown in Table 5, the performance of our approach is notably higher than the other approaches. MCT is based on computations of pixel intensity values, which suffered a significant degradation due to the poor lighting conditions as compared to the Betti Numbers approach. HOG takes both contour information and pixel intensity values, but still suffers when both features are degraded. In contrast, MSBNM and Shape Context are not influenced heavily by the pixel intensities, but was impacted due to the imperfect segmentation of the hand image due to the poor image quality. However, using topological features, our MSBNM still demonstrates better robustness than Shape Context.

| Approaches | DT | BN | CR | *Performance* |
|---|---|---|---|---|
| MSBNM (ours) | 81.3 | 73.7 | **81.4** | **81.4** |
| MCT [10] | 31.4 | 30.6 | **37.0** | **37.0** |
| HOG [17] | **65.5** | 61.7 | 65.1 | **65.5** |
| Shape Context [2] | N/A | N/A | N/A | **64.2** |

Table 5. Accuracy comparison under *cross illuminations*, wherein the training set and testing set had different lighting conditions. Shape Context is evaluated by the classifier in [2].

### 4.4. Evaluation under cross poses on various rotations

We evaluated the performance of MSBNM approach when the training set and the testing set had different hand rotations. In the first experiment, we evaluated the robustness of our approach against in-plane rotations (yaw). The yaw rotation range was divided into 4 sectors: $[0°, 90°]$, $[90°, 180°]$, $[180°, 270°]$, and $[270°, 360°]$, and the pitching/rolling of the hand was $[-30°, 30°]$ in each of the sectors above. We collected 100 samples of each posture within each rotation range, so the data set consists of 4,800 samples in total. Then, we performed a 4-fold cross-validation, in which each fold contains samples of a corresponding sector. The training set and the testing set were in different rotation ranges. In other words, the hand poses in the testing set do not appear in the training set. We refer to this experiment as *cross poses* validation. The *performance* (highest accuracy across different classifiers) is shown in Table 6.

In the second and third experiment, we evaluated the robustness of our approach against out-of-plane rotations (pitch and roll). In the training set of 1200 samples, all rotations are trivial. In the testing set of pitch, the pitching range of all 1200 test samples was $[30°, 45°]$. In the testing set of roll, the rolling range of all 1200 test samples was $[-45°, -30°] \cup [30°, 45°]$. The *performance* (highest accuracy across different classifiers) is shown in Table 6.

| Approaches | Cross pose | Pitching Test | Rolling Test |
|---|---|---|---|
| MSBNM (ours) | **86.8** | **72.4** | **67.8** |
| MCT [10] | 34.1 | 22.0 | 20.7 |
| HOG [17] | 17.9 | 46.8 | 52.6 |
| Shape Context [2] | 79.0 | 58.0 | 58.8 |

Table 6. Performance comparison wherein the training set and testing set had different rotation ranges.


Figure 11. (a) Depth images; (b) the real-time application

From Table 6, we can observe that the performance of MSBNM is significantly higher than the other three. Different rotation ranges influenced the representation of HOG and MCT features. Although Shape Context is robust to in-plane rotations, it is less robust to out-of-plane rotations than MSBNM. This verifies that our feature representation has much better rotational robustness.

### 4.5. Evaluation on a different modality

In this part, we performed the same experiment as the first experiment described in Section 3, except the images were captured using the depth camera of the Microsoft Kinect instead of a color camera. The depth camera was put in front of the hand, so the hand is segmented by using the distance between the hand and the camera as shown in Figure 11 (a).

| Approaches | DT | BN | CR | *Performance* |
|---|---|---|---|---|
| MSBNM (ours) | **89.2** | 83.0 | 88.7 | **89.2** |
| MCT [10] | 25.8 | **26.2** | 24.3 | **26.2** |
| HOG [17] | 89.4 | 81.9 | **92.1** | **92.1** |
| Shape Context [2] | N/A | N/A | N/A | **86.7** |

Table 7. Accuracy comparison for depth images. Shape Context is evaluated by the classifier in [2].

The depth images are lower quality compared to color images, and they have no texture information. Also, some part of the hand may not be captured due to the sensitivity of the camera. The accuracies of each feature representation are shown in Table 7. The accuracy of MCT is very low because MCT is based on the texture of an image. The accuracy of HOG is slightly higher than MSBNM, because our approach is more sensitive to missing parts than HOG as we discussed previously. However, the result still demonstrates the applicability of our approach to different modalities.

## 5. Application and Extension

### 5.1. Real-time application

Based on the feature representation of MSBNM, we designed a real-time application. In this application, the user can use hand postures defined in Figure 6 to interact with a program, such as drag (Posture a), draw on (Posture b), zoom in (Posture h), or reset (Posture c) the map. A snapshot of our application is shown in Figure 11 (b).

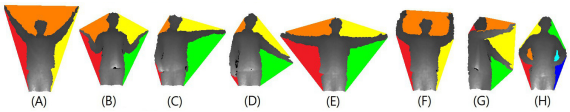


Figure 12. Eight arm postures and their complementary holes

The results demonstrate the computational efficiency of the proposed feature representation and show the feasibility for hand posture recognition in real-time.

### 5.2. Extension to arm postures

The idea of MSBNM representation of hand postures is extendible in nature to the arm postures representation as the arm postures exhibit the similar "hole" topological features under the convex hull of a body. We exploit the depth camera of Kinect to capture various arm postures. Figure 12 shows eight postures used for our study.

Using the depth camera, we captured 100 samples of each arm posture of one person as the training set, and 100 samples of each arm posture of another person as the testing data. The accuracies of each feature representation are shown in Table 8. The confusion matrix of the classification by Decision Tree using Betti Numbers is shown in Table 9.

| Approaches | DT | BN | CR | *Performance* |
|---|---|---|---|---|
| MSBNM (ours) | **93.0** | 84.0 | 89.0 | **93.0** |
| MCT [10] | 36.3 | **45.3** | 44.9 | **45.3** |
| HOG [17] | 79.6 | 50.9 | **80.0** | **80.0** |
| Shape Context [2] | N/A | N/A | N/A | **87.5** |

Table 8. Accuracy comparison of arm postures. Shape Context is evaluated by the classifier in [2].

The results are similar to the ones in Section 4.5. However, MSBNM approach yields the highest accuracy, because arm appearance is reliable during the image capture. The results in Table 8 demonstrates the applicability of using our MSBNM representation for arm posture recognition.

| A | B | C | D | E | F | G | H | |
|---|---|---|---|---|---|---|---|---|
| **64** | 0 | 0 | 27 | 0 | 9 | 0 | 0 | A |
| 0 | **98** | 0 | 0 | 0 | 0 | 0 | 2 | B |
| 0 | 0 | **100** | 0 | 0 | 0 | 0 | 0 | C |
| 0 | 0 | 2 | **98** | 0 | 0 | 0 | 0 | D |
| 0 | 1 | 0 | 0 | **99** | 0 | 0 | 0 | E |
| 9 | 0 | 0 | 5 | 0 | **85** | 0 | 1 | F |
| 0 | 0 | 0 | 0 | 0 | 0 | **100** | 0 | G |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | **100** | H |

Table 9. The confusion matrix (shown as percentages) of arm postures

## 6. Conclusion

In this paper we proposed a novel approach to analyze the topological features of hand postures at multi-scales. Since many postures do not show explicit "holes", we compute the convex hull of the hand region and consider the complementary space of the hand as holes. We use the multi-scale Betti Numbers matrix inspired by Persistent Homology to describe the multi-scale topological features of the hand posture.

Experimental results show that the multi-scale topological feature representation of hand postures by MSBNM is capable of distinguishing multiple hand postures against various illuminations, rotations, and resolutions.

In our future work, we will further analyze the Homology of the hand posture and will investigate the issue of partial occlusion by fusing it with the other texture based or contour based features.

The emerging technology of Leap Motion [12] shows very impressive results yet with very limited working range, due to the use of infrared based technique. Our approach is applicable to various modalities with a great potential to expand the working range as well as to analyze data obtained by the other devices including Leap Motion.

The multi-scale Betti numbers matrix as a new feature descriptor is also applicable for representing objects with holes or complementary holes. It, in principle, can be extended to describing any other topological objects for classification, recognition, and image information retrieval.

## 7. Acknowledgement

This material is based upon the work supported in part by AFRL and SUNY IITG. We would like to thank L. Sabalka, M. Reale, and L. Seversky for the valuable discussion.

## References

[1] V. Athitsos, J. Alon, S. Sclaroff, and G. Kollios. Boostmap: An embedding method for efficient nearest neighbor retrieval. *IEEE Trans. on PAMI*, 30(1):89–104, 2008.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape context: A new descriptor for shape matching and object recognition. *NIPS*, 2:3, 2000.

[3] J.-F. Collumeau, R. Leconge, B. Emile, and H. Laurent. Hand-gesture recognition: comparative study of global, semi-local and local approaches. *International Symposium on Image and Signal Processing and Analysis*, pages 247–252, 2011.

[4] B. Di Fabio and C. Landi. A mayer-vietoris formula for persistent homology with an application to shape recognition in the presence of occlusions. *Tech report, Universita di Bologna, Italy*, 2010.

[5] L. Ding and A. Martinez. Modelling and recognition of the linguistic components in american sign language. *Image and Vision Computing*, 27(12):1826–1844, 2009.

[6] H. Edelsbrunner and J. Harer. Persistent homology-a survey. *Contemporary Mathematics*, 453:257–282, 2008.

[7] P. Frosini and C. Landi. Persistent betti numbers for a noise tolerant shape-based approach to image retrieval. *Pattern Recognition Letters*, 34(8):863 – 872, 2013.

[8] R. Ghrist and A. Muhammad. Coverage and hole-detection in sensor networks via homology. *International Symposium on Information Processing in Sensor Networks*, 2005.

[9] K. Hu, S. Canavan, and L. Yin. Hand pointing estimation for human computer interaction based on two orthogonal-views. *ICPR*, pages 3760–3763, 2010.

[10] A. Just, Y. Rodriguez, and S. Marcel. Hand posture classification and recognition using the modified census transform. *FGR*, pages 351 –356, 2006.

[11] M. Kolsch and M. Turk. Robust hand detection. *FGR*, pages 614–619, 2004.

[12] Leap Motion, Inc. https://www.leapmotion.com/ 2013.

[13] J. MacCormick and M. Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. *ECCV*, pages 3–19, 2000.

[14] S. Mitra and T. Acharya. Gesture recognition: A survey. *IEEE Trans. on SMC, Part C*, 37(3):311 – 324, 2007.

[15] T. Moeslund and L. Norgard. A brief overview of hand gestures used in wearable human computer interfaces. *Tech Report, Aalborg University, Denmark*, 2003.

[16] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Tracking the articulated motion of two strongly interacting hands. *CVPR*, pages 1862–1869, 2012.

[17] Y. Song, D. Demirdjian, and R. Davis. Tracking body and hands for gesture recognition: Natops aircraft handling signals database. *FGR*, pages 500 –506, 2011.

[18] B. Stenger, P. Mendonca, and R. Cipolla. Model-based hand tracking using an unscented kalman filter. *Proc. British Machine Vision Conference*, pages 63–72, 2001.

[19] A. Thayananthan, B. Stenger, P. Torr, and R. Cipolla. Shape context and chamfer matching in cluttered scenes. *CVPR*, 1:I–127, 2003.

[20] J. Triesch. www.idiap.ch/resource/gestures/ Static Hand Posture Database.

[21] H. Trinh, Q. Fan, P. Gabbur, and S. Pankanti. Hand tracking by binary quadratic programming and its application to retail activity recognition. *CVPR*, pages 1902–1909, 2012.

[22] C. Vogler and D. Metaxas. A framework for recognizing the simultaneous aspects of american sign language. *CVIU*, 81(3):358–384, 2001.

[23] R. Wang and J. Popović. Real-time hand-tracking with a color glove. *ACM Trans. on Graphics*, 28:63:1–63:8, 2009.

[24] R. Yang, S. Sarkar, and B. Loeding. Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming. *IEEE Trans. on PAMI*, 32(3):462–477, 2010.

[25] Y. Yao and Y. Fu. Real-time hand pose estimation from RGB-D sensor. *ICME*, pages 705–710, 2012.

[26] H. Zhou and T. Huang. Okapi-chamfer matching for articulate object recognition. *ICCV*, 2:1026–1033, 2005.

[27] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete and Computational Geometry*, 33(2):249–274, 2005.