Matti Karppa

Estimating Hand Configurations from Sign Language Videos

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo *2014-08-08*

Thesis supervisor:

Prof. Erkki Oja

Thesis advisor:

D.Sc. (Tech.) Jorma Laaksonen

Aalto University
School of Science

AALTO UNIVERSITY
SCHOOL OF SCIENCE

ABSTRACT OF THE
MASTER'S THESIS

Author: Matti Karppa

Title: Estimating Hand Configurations from Sign Language Videos

Date: *2014-08-08* Language: English Number of pages: 101

Department of Information and Computer Science

Professorship: Computer and Information Science Code: T-61

Supervisor: Prof. Erkki Oja

Advisor: D.Sc. (Tech.) Jorma Laaksonen

A computer vision system is presented that can locate and classify the handshape from an individual sign-language video frame, using a synthetic 3D model. The system requires no training data; only phonetically-motivated descriptions of sign-language hand configuration classes are required.

Experiments were conducted with realistically low-quality sign-language video dictionary footage to test various features and metrics to fix the camera parameters of a fixed synthetic hand model to find the best match of the model to the input frame. Histogram of Oriented Gradients (HOG) features with Euclidean distance turned out to be suitable for this purpose. A novel approach, called Trimmed HOGs, with Earth Mover's Distance, as well as simplistic contours and Canny edges with the chamfer distance, also performed favorably. Minimizing the cost function built from these measures with gradient descent optimization further improved the camera parameter fitting results. Classification of images of handshapes into hand configuration classes with nearest-neighbor classifiers built around the chamfer distance between contours and Canny edges, and χ^2 distance between Pyramidal HOG descriptors turned out to yield reasonable accuracy.

Although the system displayed only moderate success rates in a full 26-class scenario, the system was able to reach nearly perfect discriminatory accuracy in a binary classification case, and up to 40 % accuracy when images from a restricted set of 12 classes were classified into six hand configuration groups. Considering that the footage used to evaluate the system was of very poor quality, with future improvements, the methods evaluated may be used as basis for a practical system for automatic annotation of sign language video corpora.

Keywords: computer vision, sign language, machine learning, metric space, computer graphics, free software

AALTO-YLIOPISTO
PERUSTIETEIDEN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Matti Karppa

Työn nimi: Käden konfiguraatioiden estimointi viittomakielisistä videoista

Päivämäärä: *2014-08-08* Kieli: Englanti Sivumäärä: 101

Tietojenkäsittelytieteen laitos

Professuuri: Informaatiotekniikka Koodi: T-61

Valvoja: Prof. Erkki Oja

Ohjaaja: TkT Jorma Laaksonen

Työssä esitetään tietokonenäköjärjestelmä, joka pystyy löytämään ja luokittelemaan käsimuotoja yksittäisistä viittomakielisten videoiden ruuduista synteettistä 3D-mallia käyttäen. Järjestelmä ei vaadi opetusdataa; pelkät foneettisesti motivoidut kuvaukset käden konfiguraatioluokista riittävät.

Kokeissa testattiin erilaisia piirteitä ja metriikoita staattisen käsimallin kameraparametrien kiinnittämiseksi, jotta löydettäisiin paras vastaavuus mallin ja syötekuvan välillä. Kokeet ajettiin realistisen heikkolaatuisella videoaineistolla. Gradienttihistogrammit euklidisella etäisyydellä osoittautuivat sopiviksi tähän tarkoitukseen. Uusi työssä esitetty lähestymistapa, jota kutsutaan trimmatuksi gradienttihistogrammiksi, maansiirtäjän etäisyyden (Earth Mover's Distance) kanssa toimi myös hyvin, kuten myös yksinkertaiset ääriviivat ja Canny-reunat chamfer-etäisyyden kanssa. Gradienttilaskeumaoptimointi (gradient descent optimization) paransi kameraparametrien sovitustuloksia. Syötekuvia luokiteltiin lähimmän naapurin luokittimilla, ja ääriviiva- ja Canny-reunapiirteiden chamfer-etäisyyteen sekä pyramidisten gradienttihistogrammien χ^2-etäisyyteen pohjautuvat luokittimet osoittautuivat toimiviksi.

Vaikka järjestelmän luokittelutarkkuus jäi vaatimattomaksi täydessä 26 luokan tapauksessa, järjestelmä saavutti liki täydellisen luokittelutarkkuuden binääriluokittelutapauksessa, ja saavutti jopa 40 % tarkkuuden, kun 12 luokan osajoukosta poimittuja kuvia luokiteltiin kuuteen eri ryhmään. Ottaen huomioon aineiston heikosta laadusta johtuvan vaativuuden, voidaan pitää uskottavana, että esitettyjä menetelmiä voidaan käyttää käytännöllisen korpusaineiston automaattiseen annotointiin soveltuvan järjestelmän pohjana.

Avainsanat: konenäkö, viittomakieli, koneoppiminen, metriset avaruudet, tietokonegrafiikka, vapaa ohjelmisto

Preface

I would like to thank Professor Erkki Oja for supervising my work, and Dr. Jorma Laaksonen for his good guidance in making this work.

I would also like to thank Dr. Son Lam Phung, for providing his segmented skin database for training the skin detector, and Dr. Leena Savolainen from the Finnish Association of the Deaf, for providing access to their dictionary video data.

Lastly, I would like to thank my wife, Tuovi, for her understanding and patience, and also for her practical help with preparing the model hand configurations.

Otaniemi, August 8, 2014

Matti Karppa

Contents

Symbols and abbreviations

Symbols

ε	Unless otherwise stated, the machine epsilon for IEEE 754 double-precision floating-point numbers (IEEE Computer Society, 2008) ($\varepsilon \approx 2.22045 \times 10^{-16}$).
h	Interval for computing difference quotient.
$\mathbf{H}$	Hidden layer output matrix (Extreme Learning Machine).
I	Image, unless otherwise noted, considered subset of $\mathbb{Z}^2$.
s	Scale parameter.
θ	Unless otheriwse stated, yaw, one of the principal axes of orientation.
t_x	Translation with respect to the x axis.
t_y	Translation with respect to the x axis.
ϕ	Unless otheriwse stated, pitch, one of the principal axes of orientation.
ψ	Unless otheriwse stated, roll, one of the principal axes of orientation.

Operators

$\|\|A\|\|_F^2$	Squared Frobenius norm, or $\sum_{i=1}^{N} \sum_{j=1}^{M} a_{ij}^2$ if a_{ij} is an element of the $N \times M$ matrix A.
$\|\|\mathbf{a}\|\|_2$	The Euclidean norm of vector $\mathbf{a}$.
$A \circ B$	The Hadamard product of matrices A and B (ie. elementwise product)
$A^{\circ x}$	Elementwise power of A (ie. each element raised to the power x)
A^+	Moore-Penrose pseudoinverse of A
$A^\intercal$	Transpose of A
$\mathcal{P}(A)$	The power set of A, ie. the set of all subsets of A: $\mathcal{P}(A) = \{B \mid B \subseteq A\}$
$c(X, Y)$	Directed chamfer distance from set X to set Y.
$C(X, Y)$	Undirected chamfer distance between sets X and Y.
$\mathrm{EMD}(P, Q, C)$	Earth Mover's Distance (EMD) between histograms P and Q, given cost matrix C
$\widehat{\mathrm{EMD}}(P, Q, C)$	*EMD hat* between histograms P and Q, given cost matrix C
$\frac{\partial f(\mathbf{x})}{\partial x_i}$	Partial derivative of function $f : \mathbb{R}^n \to \mathbb{R}$ with respect to variable x_i
$\nabla f(\mathbf{x})$	Gradient of function $f : \mathbb{R}^n \to \mathbb{R}$
$T\{I\}$	Distance transform of image I (Borgefors, 1986)

Abbreviations

AAM	Active Appearance Model
API	Application programming interface
ASL	American Sign Language
ASM	Active Shape Model
BMJ	Base metajoint
CM	Carpo-metacarpal (joint)
CPU	Central processing unit
DIP	Distal interphalangeal (joint)
DOF	Degree of Freedom
DRU	Distal radioulnar (joint)
DTW	Dynamic Time Warping
DV	Digital video
ELM	Extreme Learning Machine
EMD	Earth Mover's Distance
FN	False negative
FP	False positive
HMM	Hidden Markov Model
HOG	Histogram of Oriented Gradients
L-BFGS	Limited-memory Broyden-Fletcher-Goldfarb-Shanno
LCSS	Longest Common Subsequence
MCP	Metacarpo-phalangeal (joint)
MSE	Mean Square Error
PAL	Phase Alternating Line
PHOG	Pyramidal Histogram of Oriented Gradients
PIP	Proximal interphalangeal (joint)
POA	Place of articulation
RC	Radiocarpal (joint)
ROC	Receiver Operating Charasteristic
SIFT	Scale-invariant Feature Transform
SLFN	Single-layer Feedforward Network
SLVM	Spectral Latent Variable Model
SURF	Speeded-Up Robust Features
TGP	Twin Gaussian Process
TN	True negative
TP	True positive

1 Introduction

Hand configurations are among the most important features that convey meaning in sign languages. Other features include the place of articulation (POA), types of hand motion, and non-manual features, such as facial expressions and mouthings. Being able to classify hand configurations in sign language videos automatically is an important step for tasks such as sign language recognition and automatic annotation of corpora. Manually annotating the hand configurations in a video corpus of several hours of footage is a tedious task, and sign language researchers would find great relief in even inexact annotations when looking for video segments containing desired signs or hand configurations.

Johnson and Liddell (2011) distinguish between *hand configurations* and *handshapes*. The hand configuration is an abstract, phonetic description of the way fingers are used to articulate in a particular sign. They describe hand configurations at joint level, by assigning each joint a discrete set of different states, corresponding to minimal angles, the difference between which can be observed to affect the meaning. By handshape, they mean the perceived manifestation of a hand configuration, as produced by the signer. This terminology is used throughout this thesis.

In this work, a complete system for classifying handshapes from individual video frames into distinct hand configuration categories, characterized by a linguistically motivated abstract class description, is presented. The system works on a per-frame basis, so the loss of hand tracking information is not an issue. While the practical applicability of the present system is limited by constraints of some of its components, the general framework can remain the same while the issues are addressed with better components in future work.

The system attempts to match handshape images, given as input, to a predefined set of hand configurations. Under favorable circumstances, the system is able to locate the hands, based on skin color. The abstract hand configurations that fix joint angles are then rendered using a synthetic 3D model, and various schemes are used to fix the remaining degrees of freedom (DOF), namely the pose, or camera parameters. Finally, the handshapes are classified by computing a set features, some computationally more expensive than the ones used in the fitting stage, to discriminatively match the hand configurations to the input handshapes.

A large set of experiments were run on a set of 237 frames from Suvi (Finnish Association of the Deaf, 2003–2014) corpus, the Finnish Sign Language Web dictionary. Different feature and metric combinations were tested for probing the pose space for the best pose to match the image, including features drawn from the shape, such as the outer contour, Canny (1986) edges, and a range of descriptors based on Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005; Bosch et al., 2007). Metrics for matching these features include the chamfer distance (Barrow et al., 1977), and classical distance functions, such as the Euclidean distance and the χ^2 distance. Due to the nature of the fitting process, the primary criterion for the choice of the feature and metric combinations used herein is speed of computation. In the discriminatory stage, these were augmented by more expensive metrics, such as the Earth Mover's Distance (EMD) (Rubner et al., 1998), and descriptors

derived from Scale-invariant Feature Transform (SIFT) (Lowe, 1999) and Scaled-Up Robust Features (SURF) (Bay et al., 2006).

Since the release of Microsoft Kinect, which made an affordable depth sensor apparatus available to the public, a lot of focus has shifted towards computer vision with a depth channel. However, this work is solely concentrated on monocular RGB video, without any depth information, from Kinect, multiple cameras or otherwise. There is no doubt that the depth information would make the task easier, and offer invaluable cues, particularly when rotations out of image plane are in question. However, large video corpora already exist that have been gathered using regular video cameras, without depth sensors. While efforts are already underway to gather corpora with the help of Kinect depth sensors, pre-existing corpora also need annotations. Furthermore, linguistic information cannot be replaced by new footage because sign languages are natural languages, and as such are ever-changing. Therefore, old recordings are important from a historical point of view.

Contributions of this thesis include the description of a fully automatic system for classifying handshapes, and reports of large-scale experiments which could provide insight on what kind of features to use in future work. Particular benefits of the proposed system include the lack of requiring any training data for the given videos or individual frames, and that the proposed system is completely signer-independent. Furthermore, all components of the system are publicly available as part of the `slmotion` suite (Karppa et al., 2014).

The rest of this work is arranged as follows: Section 2 contains a brief review of previous work in the fields of hand tracking and handshape recognition. Different stages of the system, and the different features and metrics compared are presented in Section 3. In Section 4, details of the experiments are described. Results of the experiments are shown in Section 5. The results and ideas of improvement are discussed in Section 6. Section 7 concludes the thesis.

2 Previous work

Relevant previous work is presented in this section. The related work is divided into two categories: hand tracking and handshape recognition, discussed in Sections 2.1 and 2.2, respectively. Hand tracking refers to tracking the location of the hand from one video frame to another. Handshape recognition is the classification of input images of handshapes. While hand tracking may not seem quite so relevant at first sight, the work done in the field can sometimes nonetheless be applied even in recognition context, particularly when complete and autonomous systems are in question.

2.1 Hand tracking

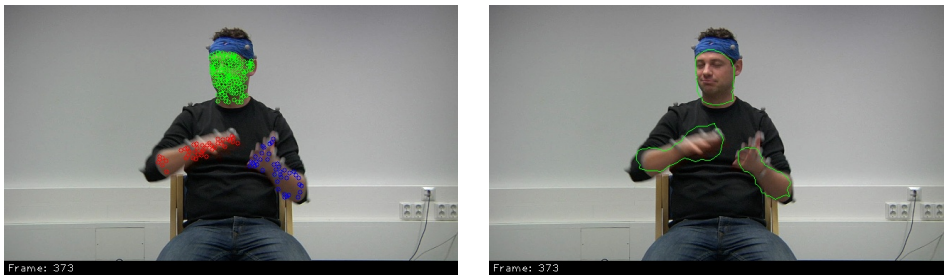
Hand tracking – following the location of the hand from one video frame to another – has been the target of computer vision research for decades. While hand tracking is not at the focus of this work, being able to locate the hand in an image in a robust fashion is such an integral part of a fully automatic hand configuration estimation system that the matter cannot be overlooked.

Early hand tracking attempts have focused on simple 2D features. Starner et al. (1998) tracked the hands by extracting a simple skin blob using a simple a priori skin model. The feature vectors they used were composed of image moments of the blob. Ambiguity between the hands was ignored, and the rightmost blob was always chosen.

Other methods built around appearance-based models include the use of Viola-Jones cascade classifier (Viola and Jones, 2001). For example, Kölsch and Turk (2004) used the Viola-Jones classifier for hand detection from individual frames. A well-known drawback of the classifier is the work required for its training. Prohibitive amounts of training data and time are required for creating a reliable signer-independent general-purpose classifier.

Karppa et al. (2011) used Active Shape Models (ASM) (Cootes et al., 1995), along with Kanade-Lucas-Tomasi (Shi and Tomasi, 1994; Lucas and Kanade, 1981) corner-point tracking based on sparse optical flow, to collect motion information. Figures 1a and 1b show the body parts as tracked with the KLT method and the ASM method. Although the method compared favorably to ground truth measured with motion capture equipment (Karppa et al., 2012), it and other simple, appearance-based methods suffer from various problems that they seem to have in common: The methods tend to handle occlusions poorly. Some approaches, such as those based on skin color, tend to place restrictions that limit their usability to laboratory settings. The head is an important place of articulation in sign languages, so an inability to locate the hand from within the area of the head is fatal. An attempt to resolve some of these issues within a skin tracking scheme was presented by Viitaniemi et al. (2013), where hand-head occlusions were detected by tracking local image neighborhoods, combined with global hand tracking. Figure 2 shows a hand-head occlusion as handled by the aforementioned method.

Buehler et al. (2008, 2011) report an arm tracking method, based on pictorial



(a) (b)

Figure 1: (a) Body parts (the hands and the face) tracked with KLT interest point tracking (Shi and Tomasi, 1994; Lucas and Kanade, 1981). (b) The contours of the skin blobs corresponding to the same body parts, tracked with ASMs (Cootes et al., 1995). The images are from (Karppa et al., 2012).

Figure 2: A hand-head occlusion as detected with the method described in (Viitaniemi et al., 2013).

structures, that could be used to locate the hand more robustly, even in the presence of occlusions. The drawback of their method is that it is not signer independent and requires a substantial amount of training data. An attempt to rectify some of these shortcomings was reported by Pfister et al. (2012), which suggests that these methods could be feasible if implemented properly.

2.2 Handshape recognition

Early work on handshape recognition has also focused on appearance-based models. Cui and Weng (2000) built a spatiotemporal, appearance-based model to handle both hand tracking and handshape recognition. Ong and Bowden (2004) classified handshapes using AdaBoost classifiers. They trained a tree of classifiers for classifying individual handshapes, on the basis of contours and edges. Wang et al. (2008) proposed a method for modeling the contour of the hand, based on Hidden Markov Models (HMM). The method is generative in nature, but, unfortunately, it is computationally very demanding. Kuzmanic and Zanchi (2007) presented a

purely appearance-based model that was based on matching images using Dynamic Time Warping (DTW) and Longest Common Subsequence (LCSS) measures.

Appearance-based approaches tend to cope poorly with out-of-plane rotations of the hand. As these are common in sign language context, this limits their usability. Discriminatory models tend to need intractable amounts of training data to cover all necessary variation in terms of pose parameters and the number of handshapes.

Recently, attention seems to have shifted towards using synthetic 3D models, as noted in Karppa (2011). Athitsos and Sclaroff (2003) modeled handshape recognition as a database retrieval problem from a pre-generated database of synthetic hand images. As a discriminatory model, the retrieval task becomes intractable when the number of handshapes grows to realistic proportions.

Stenger et al. (2006) modeled the problem as a detection problem and the configuration of the 3D model is selected via a Bayesian hierarchical detection scheme. More recently, de La Gorce et al. (2011) presented an impressive method for recovering the full hand configuration, using a sophisticated generative 3D model. The model takes into account texture continuity information and models illumination and shading. The model is fitted using quasi-Newton optimization. The authors also derived a closed-form solution to the gradient of the objective function. Despite high computational cost, and the requirement for manual initialization, the reported work is very impressive.

Very recently, when the experiments described in this work were already underway, Dilsizian et al. (2014) proposed a system based on a 3D model which was matched on the input image using Histogram of Oriented Gradients (HOG) descriptors. In this sense, their approach is very similar to the one described in this work. Their system also includes an advanced hand tracking framework. The hand tracking scheme (later called "hand location detection", see Section 3.3) presented in this work is considerably simpler and less robust. However, hand tracking was not the focus of this work.

Dilsizian et al. (2014) trained their model using data gathered from native signers using cybergloves. They train the correspondence between the HOGs extracted from their model and input images using Twin Gaussian Processes (TGP). When classifying input data, they use a Spectral Latent Variable Model (SLVM) (Kanaujia et al., 2007) for dimensionality reduction, and classify the input to one of 87 hand shapes in American Sign Language (ASL). They report recognition rates of 84 % and 71 %, with and without linguistic constraint information, respectively. Linguistic constraints relate the appearance of the hand at different frames of the sign, so, contrary to the work presented in this thesis, they also use multiple frames of information.

3 Methods

This section presents details on how the hand configuration classification system is built. Specific details are presented on the preprocessing stages the system uses, the way the remaining degrees of freedom are fixed, and finally how the classification is carried out. A lot of focus is given on the particularities of the different features and distance functions used. The level of detail given on a specific topic is strongly correlated with the amount of work required in the implementation phase: parts of the system that were implemented for this work by the author himself are covered in greater detail, including particularities of the implementation, and the parts that were readily available in third party libraries are covered only superficially.

The rest of this section is organized as follows: An overview of the entire system is given in Section 3.1. Preprocessing stages – skin and hand detection – are discussed in Sections 3.2, and 3.3, respectively. Section 3.4 presents details of the hand synthesis process. Sections 3.5–3.10 present a range of features and metrics that have been used in this work. Finally, Section 3.11 discusses the details of how to find the best matching pose for the synthetic hand.

3.1 System overview

The system works in several stages, each divided into substages. For a hand configuration recognition task at a high level of abstraction, one can identify three important steps that need to be done:

(i) Locate the hand. Hand detection is performed by first locating skin-colored areas in the input image. These areas are then labeled either as the head or a hand, based on geometric constraints and face detection data. The subimage contained within the bounding box of the skin-colored area deemed to be the hand is then selected as input for the next stage.

(ii) Find the correct scale and three-dimensional rotation of the hand models. The pose space is searched for the pose that minimizes some distance function with respect to some feature, computed for both the input image and the synthetic hand images.

(iii) Evaluate the visual similarity between the synthetic image and the input image. Once the best poses have been found, multiple features are computed and compared using several different metrics. These values are used for discriminating between different hand configuration hypotheses, given the input image.

Each particular abstract hand configuration is given as input and is fixed. Hand configuration identification can be obtained by evaluating all possible hand configurations on the same input and determining which one best corresponds to the input.

3.2 Skin detection

Within this work, skin detection means that an input color image is transformed into a binary image where the binary value of each pixel corresponds to whether the pixel should be considered skin or not. This has been achieved by classifying the pixels in the input image using an Extreme Learning Machine (ELM).

The ELM is a Single-Layer Feedforward Neural Network (SLFN), first described by Huang et al. (2006). Their description of the ELM is presented very concisely here: For an SLFN with N training samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$, $\tilde{N}$ hidden units, input weight vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{\tilde{N}}$, each N long, with biases $b_1, \ldots, b_{\tilde{N}}$, and an activation function $g : \mathbb{R} \to \mathbb{R}$, define the *hidden layer output matrix* as

$$\mathbf{H} = \begin{bmatrix} g(\mathbf{w}_1 \cdot \mathbf{x}_1 + b_1) & \cdots & g(\mathbf{w}_{\tilde{N}} \cdot \mathbf{x}_1 + b_{\tilde{N}}) \\ \vdots & \ddots & \vdots \\ g(\mathbf{w}_1 \cdot \mathbf{x}_N + b_1) & \cdots & g(\mathbf{w}_{\tilde{N}} \cdot \mathbf{x}_N + b_{\tilde{N}}) \end{bmatrix} \tag{1}$$

Furthermore, denote the output weights – vectors which map hidden units to output neurons – $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{\tilde{N}}$ and desired outputs as $\mathbf{t}_1, \ldots, \mathbf{t}_N$. With this notation, classification error can be expressed concisely as $\mathcal{E} = ||\mathbf{H}\boldsymbol{\beta} - \mathbf{T}||_F^2$ where $\boldsymbol{\beta}$ is a matrix composed of the output weight vectors $\boldsymbol{\beta}_i$ as columns, $\mathbf{T}$ is a matrix composed of desired outputs $\mathbf{t}_j$ as rows, and $|| \cdot ||_F^2$ is the squared Frobenius norm, or the sum of the square of every element.

The ELM is based on two theorems by Huang et al.: the first one states that by choosing $N = \tilde{N}$, and letting weight vectors $\mathbf{w}_i$ and biases b_j be chosen from any intervals of $\mathbb{R}^n$ and $\mathbb{R}$, respectively, $\mathbf{H}$ is always invertible and $\mathcal{E} = 0$. The second theorem states that if the activation function $g : \mathbb{R} \to \mathbb{R}$ is infinitely differentiable, then there exists $\tilde{N} \leq N$ such that $\mathcal{E} < \varepsilon$ for any $\varepsilon > 0$. Proofs of these theorems are presented in Huang et al. (2006).

One of the most important benefits of the ELM is the ease of its training. The training algorithm described by Huang et al. simply assumes fixed, randomly chosen input weight vectors $\mathbf{w}_i$ and biases b_i, which leads to the least squares solution for the output weights $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = \mathbf{H}^+\mathbf{T} \tag{2}$$

where $\mathbf{H}^+ = (\mathbf{H}^\intercal\mathbf{H})^{-1}\mathbf{H}^\intercal$ is the Moore-Penrose pseudoinverse of the hidden layer output matrix. This yields a three-step training algorithm: (i) randomly assign input weights and biases, (ii) calculate the hidden layer output matrix $\mathbf{H}$, (iii) calculate the output weights $\hat{\boldsymbol{\beta}}$.

In this particular case, the ELM is used as a binary classifier $\mathbb{Z}_{256}^3 \to \{-1, 1\}$, classifying color vectors drawn from the 24-bit RGB color space representing individual pixels in the image to either skin or non-skin classes. This is not a novel approach; a similar approach was taken, for example, in Viitaniemi et al. (2013). However, there the authors used pixels from around the face area as detected by the Viola-Jones cascade classifier as training data. In this work, another approach was adopted: a static model of the skin was created by using the data set presented in Phung et al. (2005), kindly provided by Dr. Phung of University of Wollongong.

In this particular case, the following parameters were chosen for the ELM: The activation function g was chosen to be tanh. The biases b_i were drawn uniformly from $[-3, 3]$ and the weight vectors $\mathbf{w}_i$ from $[-3, 3]^{\tilde{N}}$. The desired outputs were set to be one-dimensional with $+1$ for skin and -1 for non-skin. For choosing the number of hidden neurons $\tilde{N}$, several experiments were conducted with the data.

The data set from Phung et al. (2005), composed of 4,000 images, was used in its entirety to train the final detector. Since the total number of pixels in the data set was very high – in the order of 10^9 – a subset had to be sampled and used for training. This is because, in spite of today's computers, the amount of memory required for training the neural network with all of the data would be infeasible. This is particularly because of the need to evaluate the hidden output layer matrix $\mathbf{H}$, the size of which is $N \times \tilde{N}$ – the number of training samples multiplied by the number of hidden units.

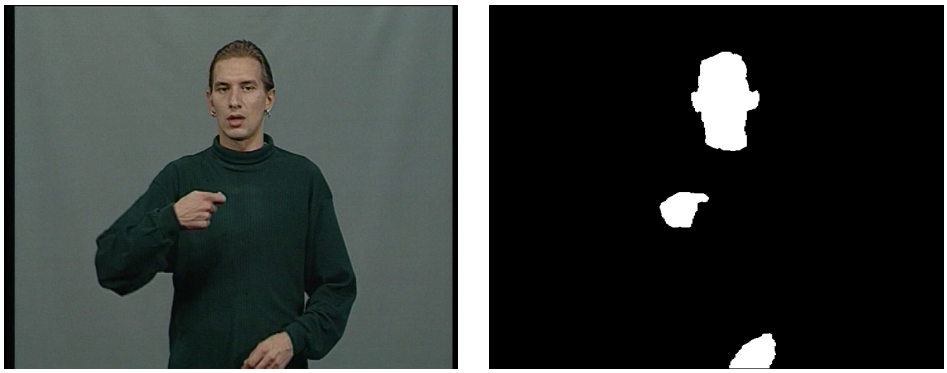
The most obvious way would be to sample the training data *uniformly*. However, in this application, positive cases outweigh negative cases in importance, and negative examples dominate the data 10:1. Therefore, another sampling scheme was devised which shall be called *balanced* sampling in this work. In that scheme, an equal number of positive and negative samples are chosen at random, subject to memory constraints.

The skin detection results are post-processed by applying first morphological opening on the binary skin mask, followed by morphological closing. Both operations are performed with a 5×5 circular kernel. This will reduce the level of noise in the resulting binary mask. The first operation erases individual misdetections and small misdetected patches. The latter operation tends to merge areas that are located very close to one another. Finally, any holes within the areas are filled by detecting contours inside the areas, and filling the polygonal areas described by the contours. An example of the end result can be seen in Figure 3. The overall quality of the results is satisfactory for the application that is at the focus of this thesis. Good performance is achieved in part due to the fact that the videos have been shot in studio conditions, there are no ambiguous elements in the background, and the signers are wearing very neutral clothes.

Performance of the skin detector was evaluated with a data set of 4,000 images (Phung et al., 2005). The authors of the data set divided the set into two equal portions, comprising 2,000 images each. In addition to original images, segmented versions of the images were provided where non-skin areas had been marked with all-white pixels. For measuring the performance of the skin detector, pixels sampled from the first 2,000 images were used as the training set, and ELMs trained with this data were used to classify the remaining 2,000 images, comprising the test set. These experiments were repeated with both uniform and balanced sampling, and with hidden neuron count $\tilde{N}$ ranging from 10 to 500.

The performance of the classifier as a function of the parameter $\tilde{N}$ was measured in terms of Precision, Recall, and Accuracy. When the classification of a pixel as skin is called a *positive* detection and as non-skin a *negative* detection, one can compute the following four values for each value of $\tilde{N}$:

- True positives (TP), the number of pixels correctly classified as skin,



(a) (b)

Figure 3: (a) The original image. (b) An example of the skin detection result.

- False positives (FP), the number of pixels incorrectly classified as skin,
- True negatives (TN), the number of pixels correctly classified as non-skin,
- False negatives (FN), the number of pixels incorrectly classified as non-skin.

Obviously, if the total number of samples classified is N, then $TP+FP+TN+FN = N$. Precision, Recall, and Accuracy are defined in terms of true and false positives and negatives as follows:

$$\text{Precision} = \frac{TP}{TP+FP}, \tag{3}$$

$$\text{Recall} = \frac{TP}{TP+FN}, \tag{4}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}. \tag{5}$$

These are standard measures for information retrieval tasks. Precision can be thought to give insight as to the fraction of retrieved objects that are relevant with respect to all objects retrieved. Recall describes the fraction of relevant objects found out of all relevant objects. Accuracy simply measures overall performance of the classifier without the notion of relevance. Although using these figures as the measure of machine learning accuracy have faced criticism (Powers, 2011), in particular because the true negative rate is completely ignored by them, they can be regarded as reasonable choices in this context since the focus is in finding skin areas accurately, while classifying non-skin is only a side effect, the performance of which is of no particular importance.

Figure 4 shows these figures for the ELM skin detector at various numbers of hidden neurons $\tilde{N}$. The different values of $\tilde{N}$ are about 50 neurons apart from one another. The figure shows how precision slightly decreases as a function of $\tilde{N}$, but recall increases noticeably, until about $\tilde{N} > 250$. In part, this may be due to the fact that the vast majority of pixels are non-skin, so at lower numbers of neurons,

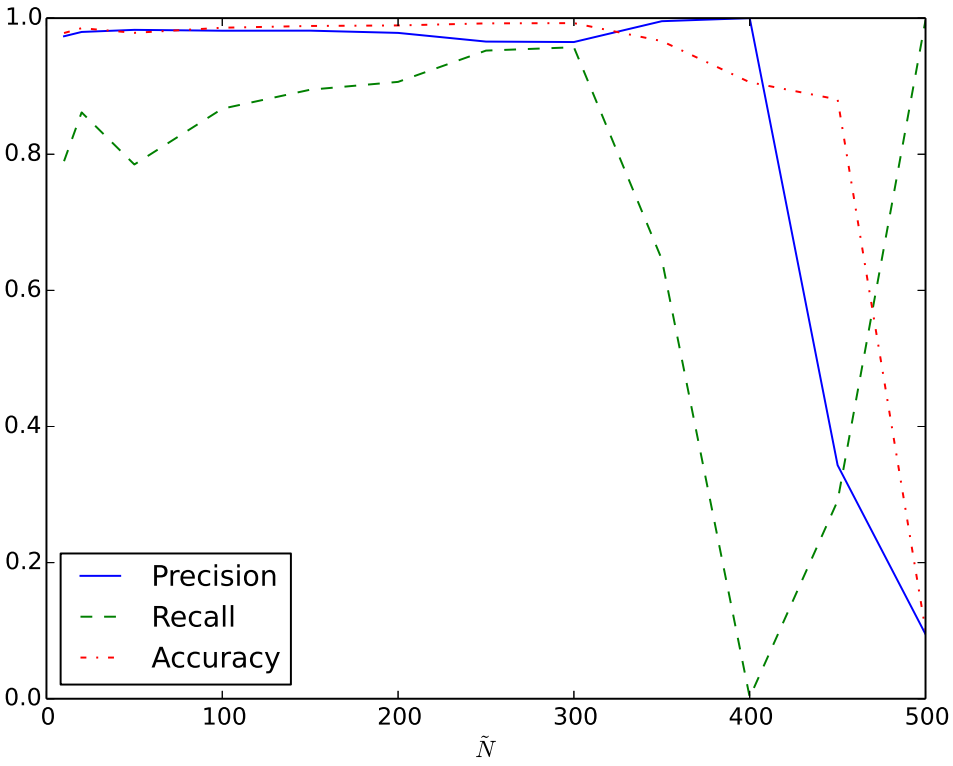


Figure 4: The ELM skin detector tested with data from Phung et al. (2005). Precision, recall, and accuracy figures are shown at various numbers of hidden neurons.

the network may tend to underestimate the number of skin pixels which favors precision. The unexpected behavior at certain numbers of neurons, such as the dent at 400 neurons may be explained by the fact that random chance plays a role in the teaching of the network, as the input weights are assigned randomly. With over 300 neurons, the behavior of the network becomes erratic. This may be because the number of training samples is limited by the memory requirement of the matrix $\mathbf{H}$ which is $\mathcal{O}(N\tilde{N})$. Even with today's computers, at a high number of neurons, fewer samples must be used for all the data to fit in the memory of the computer. This is inevitable despite the fact that experience and theory have shown that the higher the number of neurons, the greater the demand for more training data, as the number of data vectors used for training is about 5×10^8. With 500 neurons, only about 2×10^6 training vectors can be used. As a conclusion, values $200 \leq \tilde{N} \leq 300$ could be considered a good compromise, offering a reasonable accuracy and good recall while not sacrificing precision too much.

One popular way of evaluating the performance of a binary classifier is by using the Receiver Operating Charasteristic (ROC) curve. The curve measures the variation between true positive detection rate versus false positive rate as the classification threshold is adjusted. Intuitively, the curve should present the interpretation

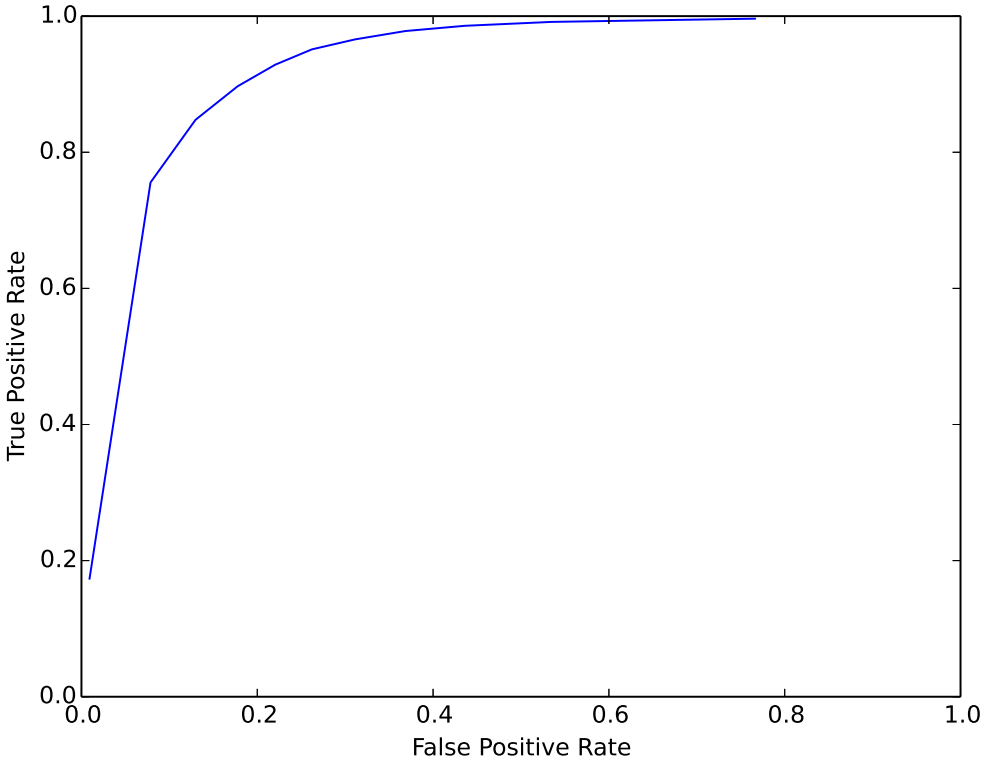


Figure 5: The ELM skin detector Receiver Operating Charasteristic (ROC) curve.

of the belief in the detections from no detections at all to interpreting all data as a positive detection. The ROC curve for the skin detector is presented in Figure 5 at $\tilde{N} = 200$ hidden neurons with the classification threshold being varied from -1.0 to $+1.0$.

3.3 Locating the hand

The hand fitting stages of the system assume that the location of the hand is known. Therefore, the location of the hand needs to be extracted in order for the system to be independent of user interaction. In this context, finding the location of the hand means determining its bounding box.

The adopted simple model bears some similarity to that of Starner et al. (1998). In principle, the hand location detection works as follows: perform skin and face detection, extract connected components from the skin mask, take the three largest components assumed to correspond to the head and the hands, remove the component which contains the face, and finally select the leftmost component, corresponding to the right hand, and compute its bounding box. All of these stages require several assumptions about the footage. Most notably, it is expected that there is only one person in the image, positioned face towards the camera. It is also assumed

that the hands do not occlude one another or the head. Furthermore, it is assumed that the signer is wearing a long-sleeved shirt. Details of these operations, apart from skin detection which was already discussed in Section 3.2, are given below.

Face detection is performed using the familiar Viola-Jones cascade classifier (Viola and Jones, 2001). The classifier, based on a cascade of weak classifiers built around Haar-like features, is trained with a boosting algorithm. In this particular case, the pre-trained classifier shipped with OpenCV (Willow Garage, 1999–2014) was used. The classifier has shown to be very robust and highly successful over the years, and seldom fails if the face is clearly visible in the image, and positioned towards the camera.

Connected components, or blobs, are extracted using the simple sequential algorithm commonly found in computer vision textbooks, first described by Rosenfeld and Pfaltz (1966). The algorithm has a linear runtime and was implemented by the author himself.

To identify the blob which corresponds to the dominant hand – assumed to be the signer's right hand – the three largest blobs were chosen. This is based on the assumption that the only skin-colored areas in the view are the face and the hands. Since the location of the face is known, it is easy to disregard the blob associated with the face, and simply choose the leftmost blob of the two blobs that remain. A bounding box is then obtained by simply iterating over the pixels, and recording the extrema. Figure 6 shows an image where the method presented here has been used for detecting the bounding box.

This methodology has several important shortcomings. There must be exactly one person, facing the camera. More importantly, the person must wear a long-sleeved shirt. This is because the synthetic hand model only covers the hand from the wrist up. In the case of a short sleeve, the hand blob would have to be cut at the wrist, which is a non-trivial task. The worst restriction, however, is the fact that the hand blobs should not be overlapped, which is the case when occlusions occur. This rules out a large fraction of possible input images. Solving these problems is a difficult and tedious task, and as such, was ruled to be beyond the scope of this work.

3.4 Hand synthesis

Handshape matching was done by the means of matching a synthetic hand image to the input image. If the terminology suggested by Johnson and Liddell (2011) is adopted when describing their phonetic model of signing, it would be misleading to talk about matching hand shapes. Rather, *hand configurations* are matched to visual handshapes.

Liddell and Johnson make a distinction between the hand configuration which is an abstract description of how a signer articulates his or her fingers, and the handshape which is the visual manifestation of the hand configuration. Liddell and Johnson provide a detailed description of discrete states for joints that they hypothesize to form phonetic minimum pairs. This is a very low-level description and the minimum changes do not necessarily affect the perceived differences at

Figure 6: Bounding box of the hand as detected by the method described in Section 3.3.

phonological level.

Matching hand configurations to input image handshapes means that an appearance of the hand is extracted from the image, and it is determined which – if any – of the abstract hand configuration descriptions appear in the image. The hand configurations were taken from Suvi, the on-line dictionary of Finnish Sign Language (Finnish Association of the Deaf, 2003–2014). To the author's best knowledge, the classification used in Suvi is the best such classification available for Finnish Sign Language, although such classifications are inherently debatable. The hand configurations were transformed into joint angle description vectors, using the 26-value parametrization described below.

The synthetic hand images were rendered using the LibHand library (Šarić, 2011). LibHand is a hand articulation library built around OGRE[1], the Open-Source 3D Graphics Engine, and OpenCV (Willow Garage, 1999–2014). The library provides a convenient way to render images of the hand by specifying joint angles and camera parameters. The author of the library, Marin Šarić, also kindly provides a

[1] `http://www.ogre3d.org/`

(a) (b)

Figure 7: Hand images rendered with LibHand (Šarić, 2011). (a) The hand in neutral position. (b) The most common hand configuration "1001" in Suvi's notation.

a

model of the hand along with the library.

Figure 7 shows examples of hands rendered with the library. The hand in Figure 7a is in "neutral" position, that is, none of the joints have been articulated from their respective 0-positions. Figure 7b shows the most common hand configuration, labeled "1001" in Suvi. Examples of other rendered hand configurations can be seen in Appendix A.

Šarić's model is composed of over 20 bones, and a total of 18 joints. Each joint has three degrees of freedom (DOF), corresponding to the three principal axes of rigid object orientation. Most of these angles are unrealistic with respect to natural human motion. Hence, the hand configurations were modeled with a subset of 25 DOFs. The choices for the four fingers and the thumb were based on observations by Johnson and Liddell (2011, 2012). The 25 DOFs are:

- For each of the four fingers:
  - The metacarpo-phalangeal joint (MCP, 2 DOFs)
  - The proximal interphalangeal joint (PIP, 1 DOF)
  - The distal interphalangeal joint (DIP, 1 DOF)
- For the thumb:
  - The carpo-metacarpal joint (CM, 2 DOFs)
  - The metacarpo-phalangeal joint (MCP, 1 DOF)
  - The distal interphalangeal joint (DIP, 1 DOF)
- The radiocarpal joint (RC, 2 DOFs)

- The base metajoint (BMJ, 2 DOFs)
- The distal radioulnar joint (DRU, 1 DOF).

With the exception of the BMJ, the nomenclature is derived directly from the common terminology used for describing the human anatomy. The joints and the bones with the associated DOFs are shown in Figure 8.

The MCP connects the metacarpi to proximal phalanges. The bony end of the proximal phalanx at the MCP end is also known as the knuckle. The proximal phalanges are connected to medial phalanges via the PIP joint, except in the case of the thumb which does not have a medial phalanx at all. The DIP joint joins the medial phalanges to the distal phalanges, and, in the case of the thumb, the proximal phalanx directly to the distal phalanx. The metacarpi and the phalanges constitute the bones of the fingers. The phalanges are the part that is visible outside, and the metacarpi are enclosed within the palm. The CM joints join the metacarpi to the carpus or the wrist. Only the CM joint of the thumb plays a significant role in hand articulation.

The RC joint joins the carpus to the radius, one of the two main bones in the forearm. The RC joint is responsible for most of what is commonly perceived as the motion of the wrist. The DRU is the joint between the radius and the ulna. The ulna is the other bone of the forearm. The DRU joint is responsible for twisting motion of the hand. The BMJ is not a true joint; it is merely the place where the hand model is cut, so the joint is entirely imaginary.

3.5 Contours and Canny Edges

The binary skin mask of the hand, as produced by methods of Section 3.2, can be used as basis for the construction of simple features describing the shape of the hand. One such feature is the outer contour of the hand. Numerous linear-time algorithms are described in the literature for extracting the contour from a binary image, and some of these algorithms are commonly included in computer vision textbooks. In this particular case, the OpenCV (Willow Garage, 1999–2014) implementation of the algorithm of Suzuki and Abe (1985) was used. Figure 9a shows an image of a hand boundary box that has been located using the method from Section 3.3. Figure 9b shows the corresponding skin mask, extracted with the ELM method from Section 3.2. Figure 9c shows the outer contour extracted from the skin mask using the algorithm of Suzuki and Abe (1985).

The Canny (1986) edge detector is perhaps the best-known general purpose edge detection algorithm. A key characteristic of the Canny detector is hysteresis: a weak candidate edge, with respect to gradient strength, is more likely labeled an edge if it is adjacent to a strong edge. The OpenCV implementation was used in this case as well. Figures 10a and 10b show an image and the extracted edges, respectively.

These two approaches, contour extraction and Canny edge detection, both produce binary images which can be treated as sets of points for the purpose of computing the chamfer distance (see Section 3.6). Being robust, well-known, and easy

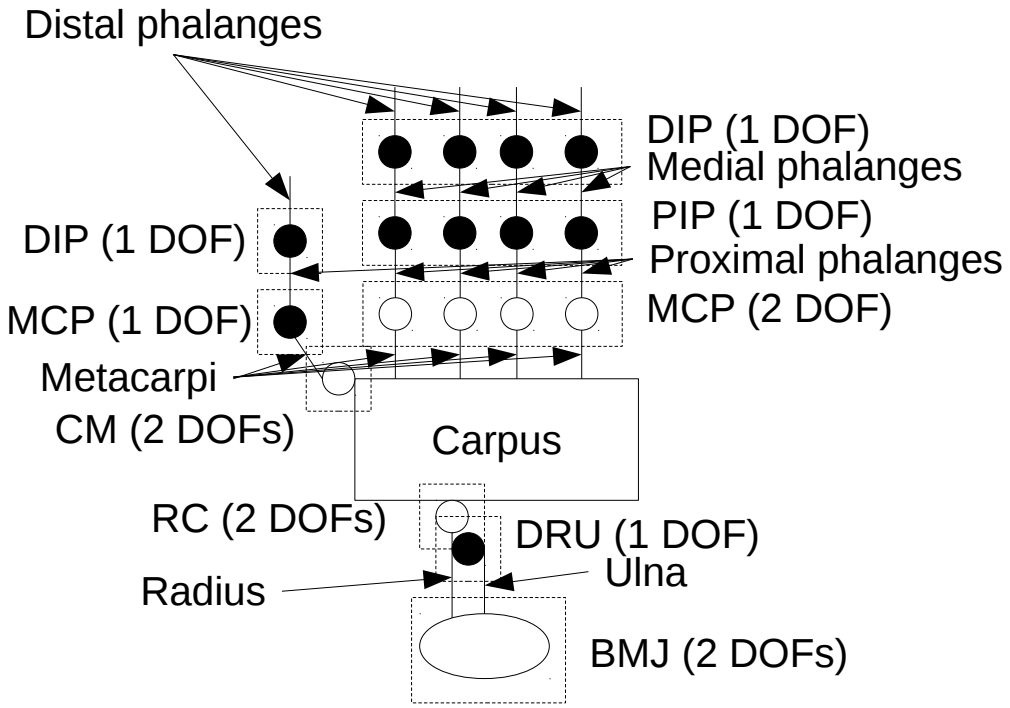


Figure 8: Bones and joints in the hand, and the 25 degrees of freedom in the reduced DOF model. Black circles denote joints with one DOF. White circles denote joints with two DOFs. In general, bones are denoted by black lines.

to compute, the point clouds produced by these two methods were chosen as baseline feature candidates: the discriminative power of a more complicated method is questionable if it cannot beat these two overly simplistic approaches.

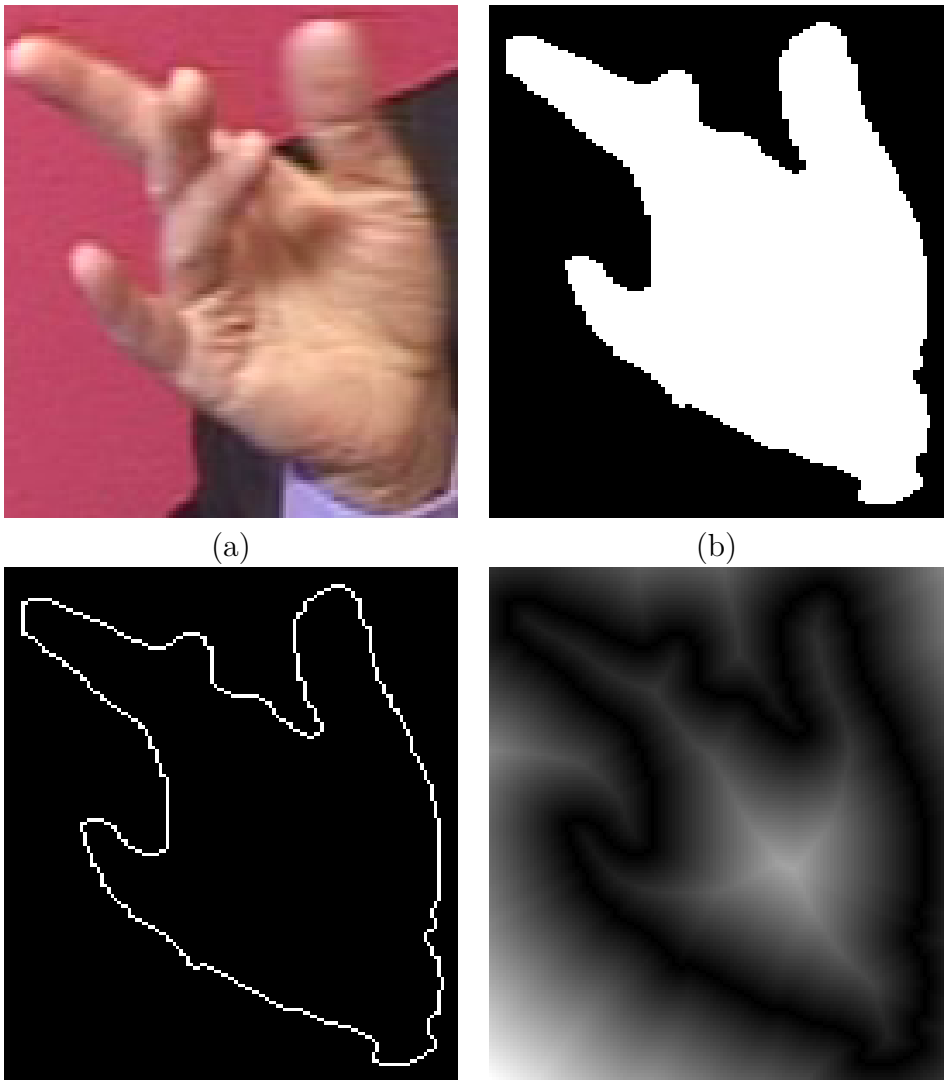
3.6 Chamfer distance

Adapting the notation from Athitsos and Sclaroff (2003), "directed" chamfer distance $c : \mathcal{P}(M) \times \mathcal{P}(M) \to \mathbb{R}^+$, from a set $X \in \mathcal{P}(M)$ to set $Y \in \mathcal{P}(M)$, both subsets of some metric space M with a ground metric d, is defined as

$$c(X, Y) = \frac{1}{|X|} \sum_{x \in X} \min_{y \in Y} d(x, y) \tag{6}$$

that is, the average shortest distance from a point in X to a point in Y. Here $\mathcal{P}(M) = \{S \mid S \subseteq M\}$ is the power set, ie. the set of all subsets of M.

The directed chamfer distance is not symmetric, and hence not a metric. A



(a) (b)

(c) (d)

Figure 9: (a) An input image. (b) The skin mask extracted via the ELM method, as described in Section 3.2. (c) The outer contour extracted using the algorithm of Suzuki and Abe (1985). (d) Distance transform of the image in (c), computed using the algorithm of Borgefors (1986).

metric can be obtained in terms of "undirected" chamfer distance

$$C(X, Y) = c(X, Y) + c(Y, X) \tag{7}$$

that is, as the sum of directed chamfer distances in both directions. For the remainder of this work, the chamfer distance may be assumed to operate on non-zero pixels in binary images, that is $M \subseteq \mathbb{Z}^2$, and the ground metric d can be assumed to be the Euclidean distance.

Application of the chamfer distance to binary images, such as contour or edge images, is very straightforward. Barrow et al. (1977) provided a linear-time algo-

(a) (b)

Figure 10: (a) An input image. (b) Edges extracted using the Canny (1986) edge detector.

rithm for computing the directed chamfer distance already in 1977. The idea is to use the distance transform: assuming that the chamfer distance is to be computed from the non-zero pixels of image I to the non-zero pixels of J, where I and J are binary images with only 1/0 values. One then defines the distance transform of image I as $T\{I\}$, an image of equal dimensions, but with real values where each value is the distance to the closest non-zero pixel. That is,

$$T\{I\}(x,y) = \min_{(i,j):I(i,j)\neq 0} d((x,y),(i,j)) \tag{8}$$

Hence, the directed chamfer distance can be computed simply as

$$c(I,J) = I \circ T\{J\} \tag{9}$$

where $\circ$ is the Hadamard product. In this particular case, the distance transform was computed with the OpenCV implementation of the algorithm of Borgefors (1986).

3.7 Histogram of Oriented Gradients (HOG)

Histograms of Oriented Gradients are descriptors used to describe the shape information contained in an image, first presented by Dalal and Triggs (2005). HOG descriptors capture both local and global variation in the image. Globality is the result of their tendency towards sparsity; it is often very easy to conglomerate descriptors of similar objects by the peaks in the histogram. On the other hand, since gradient computation is done locally by dividing the image into cells, HOGs also offer insight to the local variations exhibited by the image.

The descriptor is computed by dividing the input image into a number of *cells*. Dalal and Triggs suggested a cell size of 8×8 pixels, which has been adopted

Table 1: An example of HOG cell and block layout

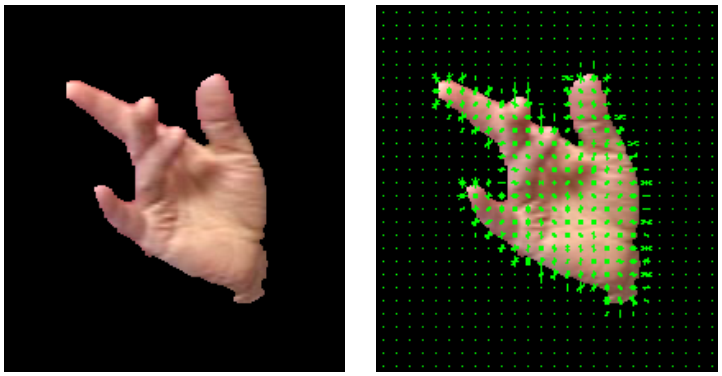
Cells					Blocks		
1	5	9	13	$\Longrightarrow$	$\{1,2,5,6\}$	$\{5,6,9,10\}$	$\{9,10,13,14\}$
2	6	10	14		$\{2,3,6,7\}$	$\{6,7,10,11\}$	$\{10,11,14,15\}$
3	7	11	15		$\{3,4,7,8\}$	$\{7,8,11,12\}$	$\{11,12,15,16\}$
4	8	12	16				

in this work. Within each cell, image gradients are computed for each pixel. A histogram of gradient orientations is formed by quantizing the gradient orientation angles into a discrete number of bins. The gradient vector, computed at each image point, then contributes an amount proportional to its magnitude to its respective orientation bin. Assuming the usual convention of selecting east (or positive x axis) as orientation 0, and b bins, the bin i would then correspond to the angle $\frac{\pi}{2} + \frac{i\pi}{b}$, where $0 \leq i < b$. Dalal and Triggs suggested a bin size of 9 different orientations. They also suggested that simple one-dimensional kernels $[-1, 0, 1]$ and $[-1, 0, 1]^\intercal$ be used for gradient computation. Following the advice, these choices have been adopted in this work.

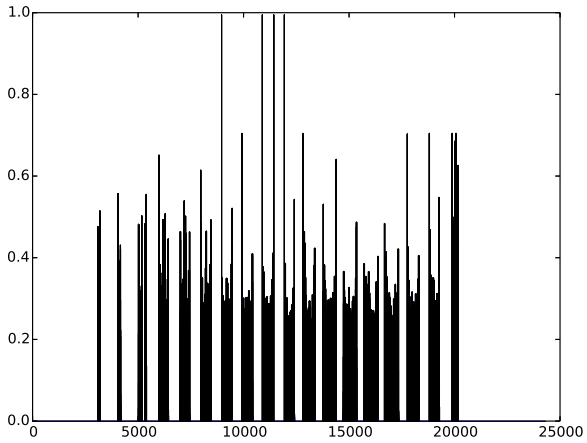
Cells are then grouped into partially overlapping *blocks*. Following suggestions given by Dalal and Triggs (2005), blocks consisting of 2×2 cells were adopted. In this case, partial overlap means that horizontally or vertically adjacent blocks have two cells in common, and diagonally adjacent blocks have exactly one cell in common. As a practical example, consider Table 1. There each number corresponds to a cell number. On the right hand side, the cells have been grouped into blocks. As an implementation detail, it should be noted that the block layout follows a column-major order. Also, it should be noted that, with the exception of the corner cells, each cell contributes to at least two blocks. Although Dalal and Triggs suggested that, as a final step, the values of each block should be normalized to sum up to unity, this block-normalization has not been done within this work for the vanilla HOG.

The descriptor is formed as a concatenation of blocks, where each block consists of a concatenation of its cells, and each cell corresponds to a gradient histogram. Assuming n blocks, with c cells each, and b bins, this yields an ncb-long descriptor. With the choices above, for images in the order of 200×200 pixels, this yields a descriptor length on the order of 2×10^4, which is rather long. On the other hand, the histograms are often sparse, which allows for easy size reduction through removal of zero or otherwise very small components.

As a practical example, Figure 11a shows a 208×224 image of a hand where non-skin pixels have been blacked out. Figure 11b shows the HOG structure: at each cell, the histogram values are shown superimposed on the image as vectors, corresponding to the orientation of the bin, and with scale corresponding to the respective histogram value. While the block structure is not visible here, the image shows how the HOG descriptor captures local variations. Also, the values have meaningful interpretations, considering the appearance of the image. Figure 11c



(a) (b)



(c)

Figure 11: (a) Input image of a hand with non-skin pixels blacked out. (b) Gradient values drawn int the middle of cells. (c) The corresponding HOG histogram. No scaling has been done, resulting in blocks of variable size.

shows the actual histogram. The histogram exhibits typical behavior, such as a limited number of very strong peaks and gaps corresponding to all-black values.

Pyramidal Histogram of Oriented Gradients (PHOG). Bosch et al. (2007) presented a pyramidal extension to the HOG descriptor. The descriptor they proposed works by computing the gradient histograms at multiple levels. On the first level, the histogram is computed for the entire image. In subsequent levels, every cell in the image is divided into four cells of equal size, and the gradient computation is performed recursively in each cell. The final descriptor is obtained as a concatenation of all these sub-histograms. With L levels and B bins, this yields a descriptor of N real values where

$$N = B \sum_{l=1}^{L} 4^{l-1} = B \frac{4^L - 1}{3}. \tag{10}$$

Algorithm 1 Computing the PHOG by utilizing an existing HOG implementation.

Given access to an existing HOG implementation, computes the PHOG descriptor of input image I at L levels. Let $||$ denote vector concatenation, and let $\mathrm{HOG}(I, b, c, s, B)$ be the HOG descriptor of image I with block size $b \in \mathbb{N}^2$, cell size $c \in \mathbb{N}^2$, block stride $s \in \mathbb{N}^2$, and $B \in \mathbb{N}$ bins, with b an integer multiple of c.
function PHOG(I: an $n \times m$ input image, B: the number of orientation bins, L: the number of levels in the pyramid)
    Initialize PHOG vector P as empty.
    for $l := 0$ to $L - 1$ **do**
        Compute $H := \mathrm{HOG}(I, n \times m, \frac{n}{2^l} \times \frac{m}{2^l}, \frac{n}{2^l} \times \frac{m}{2^l}, B)$.
        Set $P := P||H$.
    end for
    return P
end function

This approach gives flexibility, as the size of the descriptor is easy to adjust by an appropriate choice of the number of levels. At the same time, the descriptor truly provides both global and localized description. Furthermore, an existing HOG implementation can be used to compute the PHOG descriptor by evaluating the HOG with only one block and a power of 4 number of cells at each level, concatenating the resulting descriptors. This procedure is outlined in Algorithm 1. Following Bosch et al.'s advice, the resulting histograms are normalized to sum up to unity.

While Bosch et al. gave no definite suggestions as to the number of levels or bins, the number of bins used in this work, $B = 20$, was chosen based on their experience. The experiments were conducted with $L = 3$, 4, and 5 levels, yielding 21, 85, and 341 value descriptors, respectively. Bosch et al. also suggested that the χ^2 distance be used for comparing two PHOG descriptors. The χ^2 distance will be described in Section 3.9.

Figure 12 shows the first four levels of a PHOG histogram, computed at five levels in total. The cell division is presented at each level, followed by the actual histogram. It is easy to see that each successive level provides more localized and detailed information about the image content, as expected.

HOG bin distance. Later on in this work, distance functions are presented that are parametrized with respect to other distance functions, such as the Earth Mover's Distance described in Section 3.10. These metrics passed as arguments will be referred to as ground metrics. This gives motivation to the construction of a ground metric between HOG bins in order to construct such a high-level metric between different HOG descriptors. A novel simple distance function will be described and justified next.

Each bin in the HOG histogram is associated with a pair $(\mathbf{p}, \theta)$, where $\mathbf{p}$ is a vector describing the location where the gradients were extracted, and θ describes the direction of the gradients that were classified into the given bin. $\mathbf{p}$ shall be assumed to be equal to the centroid of the corresponding cell.

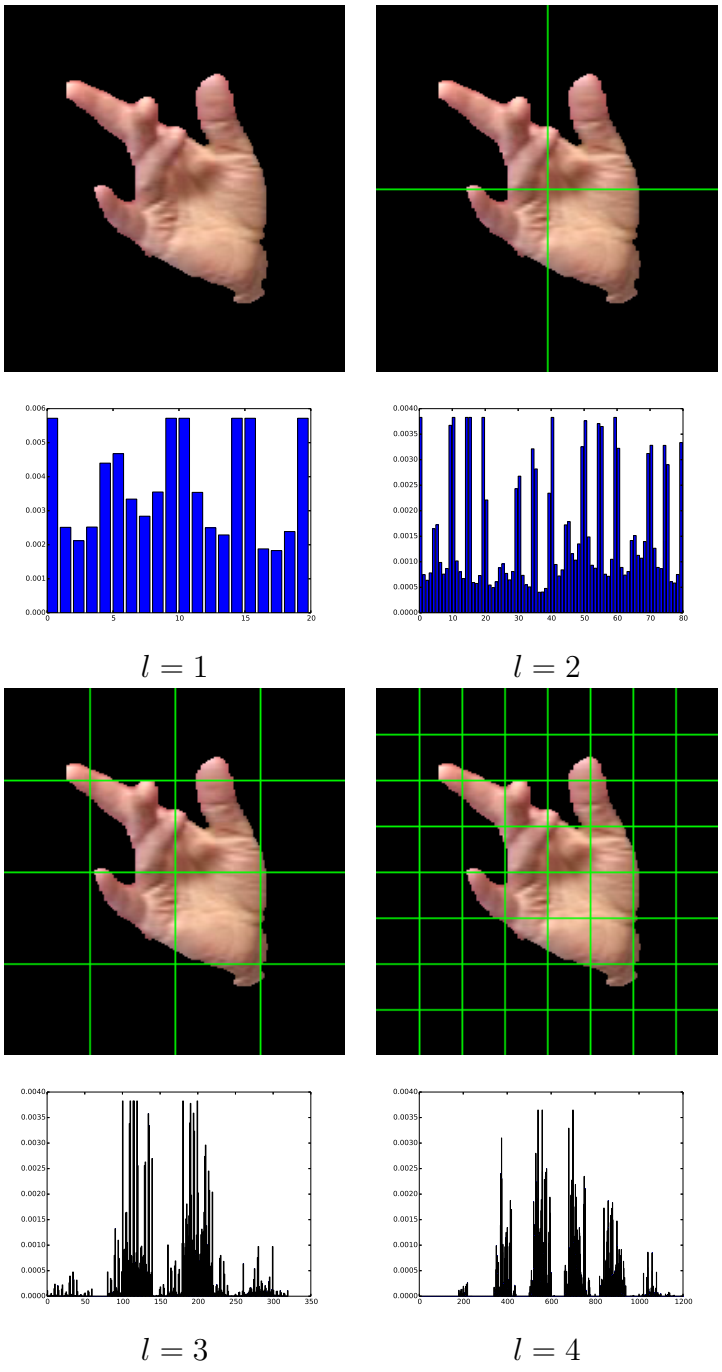


Figure 12: Example of the first four levels of a PHOG with five levels in total. The upper image shows the cells, and the lower image shows the actual histogram.

Intuitively, it would be appealing if the distance between two bins would be dominated by the distance between the corresponding cells if the cells are far apart, and dominated by the angular difference if they are close by. Given two bins i and j, associated with the pairs $(\mathbf{p}, \theta)$ and $(\mathbf{q}, \phi)$, respectively, with $0 \leq \theta, \phi < \pi$, define the distance between the two bins as the sum of the Euclidean distance between

the cell centroids, and a weighted absolute difference of the angular distance. This yields

$$d(i,j) = ||\mathbf{p} - \mathbf{q}|| + w|\theta - \phi|. \tag{11}$$

As the HOG bin distance is defined as a sum of L_1 and L_2 metrics, it is trivially metric by itself. Unless otherwise stated, w was set arbitrarily to correspond to the cell width within this work. Unfortunately, due to lack of time, it was deemed to be beyond the scope of this work to conduct experiments with the behavior of the metric at different values of w, and with other ways of handling the angular differences that would not violate the triangle inequality.

Trimmed HOG. One problem with the HOG descriptors is that they tend to be very long. For the hand images in the experiments of Section 4, the size of a typical HOG descriptor would be on the order of 10^4 floating-point values. As it may be necessary to construct distance matrices between HOG bins, for example by using the bin distance described in Equation (11), using the entire HOG to do this would yield a distance matrix with a size in the order of hundreds of megabytes, or even gigabytes, if floating-point numbers are used. This is very impractical, especially since, as noted earlier in this section, the descriptors tend to be somewhat sparse, and most bins have little to contribute to the overall description.

As a novel solution to this problem, let H be a HOG descriptor of n values, whose values are assumed to be unnormalized with respect to blocks. Reorder the bins into a descending order with respect to the value of the bin. Denote this reordered descriptor by H'. Then, create a cumulative histogram C such that

$$C_i = \frac{\sum_{j=1}^{i} H'_i}{\sum_{j=1}^{n} H'_i} \tag{12}$$

so each component in C describes the *proportion* of the i most weighted values in the HOG descriptor of the total mass of the descriptor.

From here on, there are two ways to proceed. It is possible to determine a threshold value $t \in [0, 1]$, and drop any components which do not fall into the most significant t fraction of the histogram. Alternatively, a simple maximum number of the most significant values can be taken. The resulting descriptor is then called a *trimmed* HOG in this work.

For example, Figure 13 shows two skin-segmented hand images whose HOG descriptors have been computed in this fashion. The sizes of the descriptors for images in Figure 13a and 13b are 24,300 and 64,664, respectively. Figure 14 shows the cumulative histograms for these HOGs. The first 500 most weighted components contain 20 % and 10 % of the total mass of their respective histograms, respectively. 99 % of the total mass is contained in the 5,765 and 14,305 most weighted values, respectively. This shows that a large fraction of the bins can be discarded without losing too much information.

(a) (b)

Figure 13: Input images that were used for computing exemplary trimmed HOGs.

3.8 Local features: SIFT and SURF

Scale-invariant Feature Transform (SIFT) (Lowe, 1999) and Speeded-Up Robust Features (SURF) (Bay et al., 2006) are two related algorithms that find features or interest points with favorable qualities. Another key feature is that they also provide concise descriptors describing the qualities of the feature points found, which can be used for interest point matching. Both of these algorithms have seen wide use in object recognition.

SIFT is the older one of the two algorithms, first described by Lowe (1999). As the name suggests, the algorithm is invariant to scaling and also to changes in orientation. Key points are located as maxima and minima of differences of Gaussians, applied to suitably processed images in scale space. The detected key points are described by a vector of 128 real values.

SURF was described by Bay et al. (2006), and is claimed to be more robust than SIFT. SURF key points are detected using 2D Haar wavelet responses, and are described by vectors of 64 real values.

The OpenCV (Willow Garage, 1999–2014) library contains an implementation of both of these algorithms, and these implementations were used in these experiments.

3.9 Simple metrics: The Euclidean norm, and the χ^2 distance

The Euclidean norm is possibly the most widely known distance function between two vectors of equal length. Familiarly, the Euclidean distance between two vectors $P, Q \subseteq \mathbb{R}^n$ is defined as

$$d_e(P,Q) = \sqrt{\sum_{i=1}^{n}(P_i - Q_i)^2}. \tag{13}$$

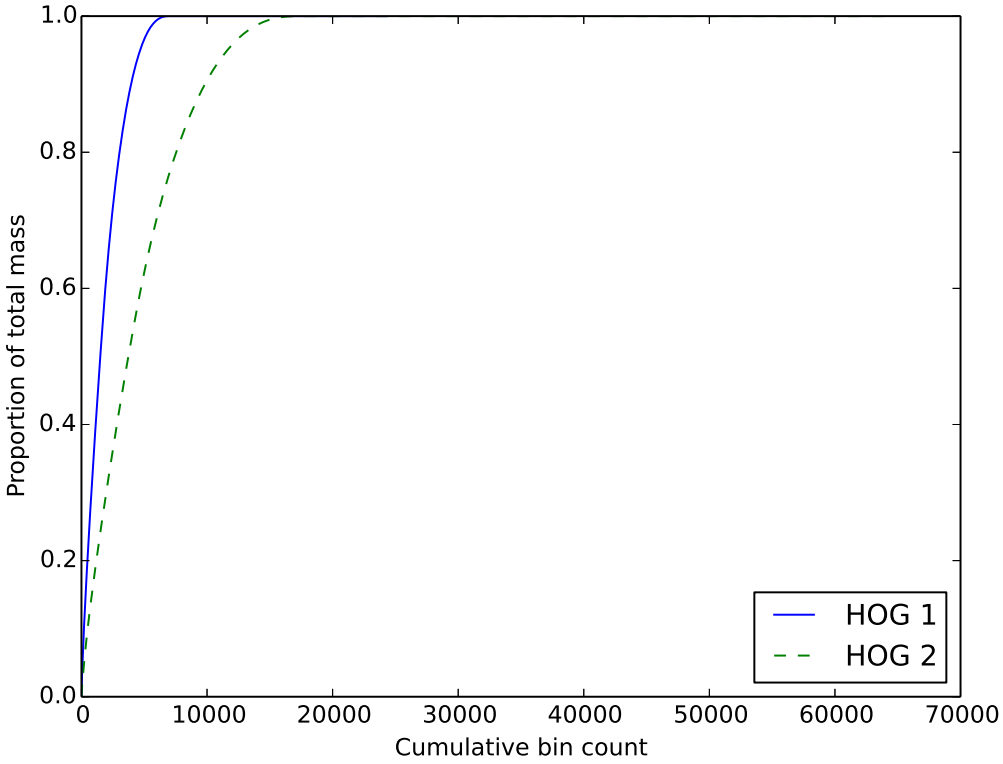


Figure 14: Cumulative histograms for HOGs computed for images in Figures 13a and 13b. The sizes of the descriptors were 24,300 and 64,664, respectively, with all of the mass contained by 5,765 and 14,305 bins, respectively.

Applying the Euclidean distance between histograms is simply a matter of treating them as vectors.

Another well-known simple distance function between equally-long histograms is the χ^2 distance, inspired by the Pearson's χ^2 test from statistics. The χ^2 distance is defined as

$$d_{\chi^2}(P,Q) = \frac{1}{2}\sum_{i=1}^{n}\frac{(P_i - Q_i)^2}{P_i + Q_i}. \tag{14}$$

What both of these metrics have in common is that they are very easy and fast to compute, and have intuitive interpretations. The fastness of the computation is even more obvious if Equations (13) and (14) are restated in terms of matrix operations, assuming P and Q are row-vectors:

$$d_e(P,Q) = \sqrt{(P-Q)(P-Q)^\intercal} \tag{15}$$

$$d_{\chi^2}(P,Q) = \frac{1}{2}(P-Q)^{\circ 2}\left((P+Q)^{\circ -1}\right)^\intercal \tag{16}$$

where $\circ$ denotes Hadamard, that is elementwise, powers.

Downsides of these metrics include the fact that, when applied to histograms, as pointed out by Pele and Werman (2008), there must be a very good agreement between the bins from the different histograms. A small error in alignment can have devastating effects if datapoints that correspond to one another in the intended sense systematically fall into different-numbered bins.

3.10 Earth Mover's Distance (EMD)

The Earth Mover's Distance (EMD) is a metric between distributions, such as histograms. While the idea itself is age-old, related to the transportation problem, the name EMD was proposed by Rubner et al. (1998). What follows is a brief overview of the definition of the EMD, an extended definition, and a brief description of how it is computed. A more complete treatment of the EMD can be found in the paper by Rubner et al. (1998).

Let P and Q be two histograms with each bin P_i, Q_j associated with an element from a metric space $\mathcal{M}$. In principle, one could consider the histograms to be weighted point clouds. The histograms shall *not* be assumed to be normalized, nor do they need to sum up to unity. Let the number of bins in histograms P and Q be N and M, respectively, so $P \in \mathbb{R}^N$ and $Q \in \mathbb{R}^M$. Suppose $p_i \in \mathcal{M}$ is the point in the metric space that corresponds to bin i in P, and $q_j \in \mathcal{M}$ for the bin j in Q. Let $d' : \mathcal{M} \times \mathcal{M} \to \mathbb{R}^+$ be the distance function associated with the space $\mathcal{M}$. Define $d : \mathbb{Z}_N \times \mathbb{Z}_M \to \mathbb{R}^+$ as $d(i,j) = d'(p_i, q_j)$, and let us call d *the ground metric* from now on. The EMD will be defined with respect to the ground metric.

It can be thought that P represents a mass of *earth* or dirt that is used to fill the holes, represented by Q. Assuming that the mass of P is greater than or equal to that of Q, that is, $\sum_{i=1}^{N} P_i \geq \sum_{j=1}^{M} Q_i$, the EMD seeks to minimize the amount of work needed to transport the dirt P to holes Q so that all the holes get filled. The work is considered proportional to the distance between piles of dirt and holes. Formally, denoting $c_{ij} = d(i,j)$, this would be equal to minimizing the total cost of the flow f_{ij}:

$$d(P,Q) = \min_{\{f_{ij}\}} \sum_{i=1}^{N} \sum_{j=1}^{M} c_{ij} f_{ij} \tag{17}$$

subject to

$$f_{ij} \geq 0 \; \forall i,j \tag{18}$$

$$\sum_{j=1}^{M} f_{ij} \leq P_i \; \forall i \in \{1,\ldots,N\} \tag{19}$$

$$\sum_{i=1}^{N} f_{ij} = Q_j \; \forall j \in \{1,\ldots,M\}. \tag{20}$$

that is, all holes are filled and no more dirt is used than is available. It can be shown that when d is a true metric and $\sum_{i=1}^{N} P_i = \sum_{j=1}^{M} Q_j$, then the EMD is a true metric as well.

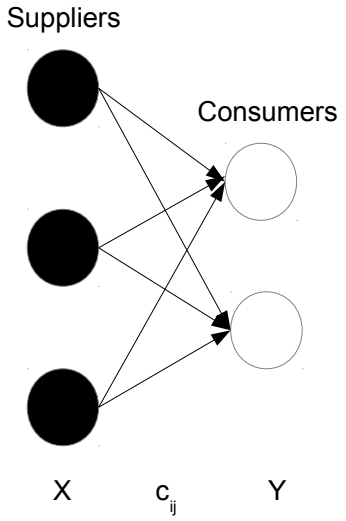


Figure 15: A bipartite graph representing the EMD as a transportation problem from suppliers to consumers. This figure is a recreation of Figure 1 of Rubner et al. (1998).

One way to view the problem is as a transportation problem by estimating the optimal flow of goods from *suppliers* to *consumers*. This yields an instance of the max-flow-min-cost problem in a bipartite graph, such as the one in Figure 15. Once the optimal flow has been found, the EMD is defined as

$$\mathrm{EMD}(P, Q, C) = \min_{\{f_{ij}\}} \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} c_{ij} f_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{M} f_{ij}} = \min_{\{f_{ij}\}} \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} c_{ij} f_{ij}}{\sum_{j=1}^{M} Q_j} \tag{21}$$

where the denominator is a normalization factor to avoid smaller total weights from being favored, and C is the $N \times M$ cost matrix with elements $c_{ij} = d(i, j)$.

Rubner et al. (1998) suggest that the transportation Simplex algorithm should be used for finding the optimal flow, and provide a working C implementation[2]. A comparison between certain other implementations will be shown later in this section.

The EMD has several advantages compared to some other metrics between histograms. Compared to the L_2 or Euclidean metric, or the χ^2 metric, the EMD does not need such strict assumptions about normalization of the bins. For the former two metrics to be meaningful, the number of bins between the histograms must be equal, and each bin must have precisely the same interpretation. Otherwise, it is not possible to compute a meaningful distance; bin number requirement is a mathematical necessity, and if the histograms have not been normalized very carefully, a comparison between apples and oranges may ensue.

[2] `http://robotics.stanford.edu/~rubner/emd/default.htm`

The EMD is not perfect, however. First and foremost, computing it may be very costly if the histograms are large. In fact, the transportation Simplex has an exponential worst-case runtime, although this situation very rarely occurs. If the histograms are sparse, a lot of bins may be dismissed to reduce the runtime, but this may not always be the case. Another problem is the fact that the EMD is not guaranteed to be a metric if the histograms have inequal masses. In some cases, the difference in the masses may be a significant cue for the discrimination between different classes of objects. Also, interpreting the EMD as the amount of *energy* required to transform one distribution into another, it can be a problem if the extra mass is simply ignored.

To rectify the latter problem, Pele and Werman (2008) have proposed an extension to the EMD which they have dubbed $\widehat{\text{EMD}}$ (pronounced *EMD hat*). The difference to the vanilla EMD is that $\widehat{\text{EMD}}$ is not normalized, and it penalizes the extra mass of the larger histogram. It is defined as

$$\widehat{\text{EMD}}(P,Q,C) = \min_{\{f_{ij}\}} \sum_{i=1}^{N}\sum_{j=1}^{M} f_{ij}c_{ij} + \left|\sum_{i=1}^{N} P_i - \sum_{j=1}^{M} Q_j\right| \alpha \max_{ij} c_{ij} \tag{22}$$

where α is a variable that controls the amount of penalty imposed for extra mass. For $\alpha = 0$, $\widehat{\text{EMD}}(P,Q,C) = \text{EMD}(P,Q,C)$. Pele and Werman showed that $\widehat{\text{EMD}}$ is guaranteed to be a metric for $\alpha \geq 0.5$. An existing EMD implementation can be used to compute the $\widehat{\text{EMD}}$ by the rewriting Equation (22) as follows:

$$\widehat{\text{EMD}}(P,Q,C) = \text{EMD}(P,Q,C) \min\left(\sum_{i=1}^{N} P_i, \sum_{j=1}^{M} Q_j\right) + \left|\sum_{i=1}^{N} P_i - \sum_{j=1}^{M} Q_j\right| \alpha \max_{ij} c_{ij}. \tag{23}$$

Since implementations of the EMD were readily available, it was deemed unnecessary to implement one anew. Three different EMD implementations were compared:

- Rubner's original C implementation of the EMD, based on transport Simplex (Rubner et al., 1998)
- A similar C++ implementation found in the OpenCV library (Willow Garage, 1999–2014)
- Pele and Werman's C++ implementation of their *FastEMD* (Pele and Werman, 2009).

The first two implement the vanilla EMD, while the last one implements $\widehat{\text{EMD}}$. The last implementation also places certain limitations: first of all, it requires that both histograms have an equal number of bins. Secondly, the cost matrix must be symmetric and positive semidefinite. This means that, if interpreting the histograms as representing weighted point clouds, the two histograms must be drawn from clouds with equal points, but possibly different weights.

Taking the aforementioned limitations into account, the performance of these different implementations was compared in two sets of synthetic tests. At the same

time, it was verified that all three implementations agreed on the solution. The first test studied the vanilla EMD performance between two N-bin-histograms where each bin i corresponds to a randomly generated point (x_i, y_i) where x_i, y_i have been drawn uniformly from $[-100, 100]$. The Euclidean distance was chosen as the ground metric. The two histograms were generated by drawing $2N$ numbers uniformly from $[0, 1]$. The histograms were then normalized to sum to unity.

In the second test, the OpenCV implementation was compared to Pele and Werman's implementation by computing the $\widehat{\text{EMD}}$ with $\alpha = 1.0$. In the case of OpenCV, this was achieved by the formulation of Equation (23). The cost matrix and the histograms were generated as above, with the exception that the histograms were not normalized.

Results of the first and second test are shown in Figures 16, and 17, respectively. Figure 16 shows that there is a substantial constant term in the runtimes of Rubner's C implementation. This is likely related to the fact that Rubner's code does not admit a precomputed cost matrix, but takes in a distance function, which it uses to populate a cost matrix whose size is hard-coded. Pele and Werman's implementation did not perform very well for large histogram sizes. This may in part be due to the fact that, unlike they suggested (Pele and Werman, 2009), the ground metric was not thresholded. Asymptotically, disregarding the constant term, the OpenCV implementation performed very similarly to that of Rubner's, which was to be expected since they both share the transport Simplex approach to the problem. The results for computing the $\widehat{\text{EMD}}$ in Figure 17 verify these observations, and lead to the conclusion that the OpenCV implementation is the best choice for the task at hand. As a sidenote, it turned out that, unlike specified by Pele and Werman (2008) and Rubner et al. (1998), in Pele and Werman's implementation, the EMD was normalized by dividing the value by the *maximum* of the two histogram sums, rather than the minimum. While very straightforward, it takes two runs of the algorithm to address this issue, as the algorithm needs to be run with and without extra mass penalty, if the correctly scaled value is sought.

3.11 Determining the best pose

When the hand configuration is fixed, there are six degrees of freedom that ultimately determine the appearance of the rendered hand. These are:

- The scale s
- The three principal axes of orientation θ, ϕ, and ψ (or yaw, pitch, and roll)
- The two translation parameters t_x and t_y.

As the signer is always assumed to be positioned frontally, the *yaw* axis can be taken to be in the range from 0 to 180° degrees; the other two angles retain the full range, however. As the translation parameters t_x and t_y tend to be very close to zero, they are not taken in account when the initial pose is probed. They may only be updated to be non-zero if one of the iterative optimization schemes, described later, so sees

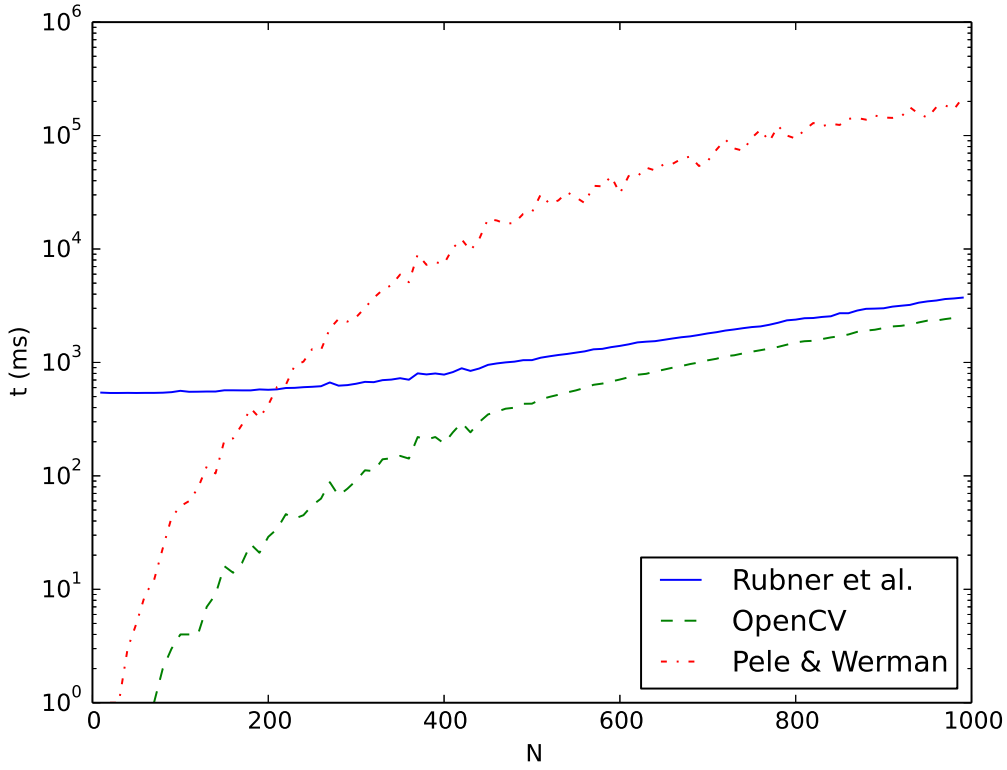


Figure 16: Runtimes of different EMD implementations as function of histogram size N. Note that the scale of the vertical axis is logarithmic.

fit. The initial scale parameter was estimated by selecting it in such a manner that the rendered hand just fits within the bounding box.

All feature and metric combinations described above tend to produce cost functions that have numerous local minima. This means that simple optimization methods tend to be very sensitive to the choice of the initial pose. This makes it necessary to probe multiple initial hypotheses.

Five different schemes were attempted. Three of these include simple probing of the three-dimensional orientation space at different steps: 45°, 30°, and 22.5°. In addition to this, two different optimization schemes were tried: the gradient descent and the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) optimization (Byrd et al., 1995). In both cases, the initial pose space is explored at 45° intervals, and the hypothesis is then refined using the optimization scheme.

Both optimization methods require the gradient of the cost function to be computed. A general-purpose solution was implemented by the means of the simple difference quotient approach:

$$\frac{\partial f(\mathbf{x})}{\partial x_i} \approx \frac{f(x_1, \ldots, x_i + h, \ldots, x_n) - f(x_1, \ldots, x_i - h, \ldots, x_n)}{2h}. \tag{24}$$

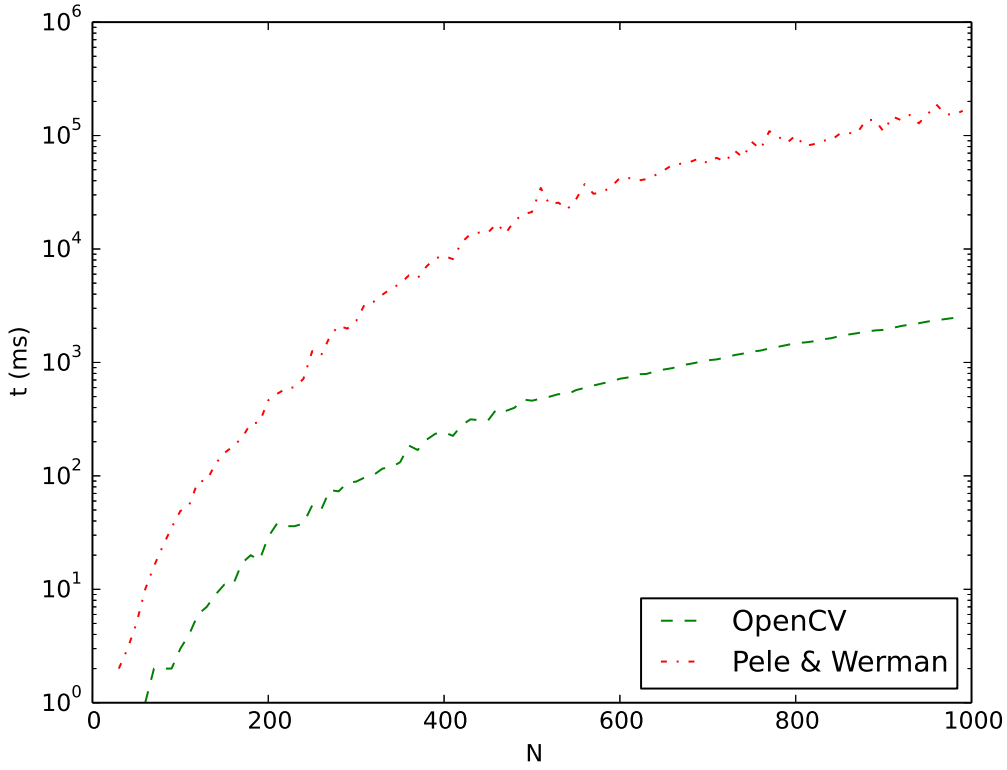


Figure 17: Runtimes of different $\widehat{\text{EMD}}$ implementations as function of histogram size N. Note that the scale of the vertical axis is logarithmic.

Press et al. (2007, p. 230) suggested that the interval h should be set to $h = \Delta x_i = x_C \sqrt[3]{\varepsilon}$ where $\varepsilon \approx 2.22045 \times 10^{-16}$ is the machine epsilon for IEEE 754 double-precision floating-point numbers (IEEE Computer Society, 2008), and $x_C \equiv \sqrt{f/f''}$ is the curvature of the function. They also suggest that, lacking other information, $x_C = x_i$ can be chosen. However, about zero, the estimate goes to zero, so the following interval is adopted instead to prevent this behavior:

$$h = \max(|x_i| \sqrt[3]{\varepsilon}, \sqrt[3]{\varepsilon}). \tag{25}$$

The cost function is evaluated by re-rendering the hand with the new pose, and recomputing the features that are used for evaluation. This raises suspicions that the chosen interval may be too short in some cases, as there is bound to be some limit as to how small changes can actually be seen in the new rendered images. However, determining this was deemed to be beyond the scope of this work.

The gradient descent is a classic optimization algorithm. Given initial parameter values $\mathbf{x}_0$ and the cost function f, let $\mathbf{x}_i$ be the parameter vector at iteration i. The update rule for minimizing $f(x)$ is

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \gamma_i \nabla f(\mathbf{x}_i) \tag{26}$$

where $\gamma_i = 2^{-k}$ and $k \in \mathbb{Z}^+ \cup \{0\}$ is the minimal integer, such that

$$f(\mathbf{x}_{i+1}) < f(\mathbf{x}_i). \tag{27}$$

That is, the parameter γ_i is adjusted to prevent overshooting. Optimization ends when $\nabla f(\mathbf{x}_i) = \mathbf{0}$, no new γ_i can be determined, or the maximum number of iterations is exceeded. Due to lack of precision, and the discrete nature of the cost function, it can happen that the first condition is never reached.

The L-BFGS (Byrd et al., 1995) is an optimization algorithm belonging to a class of algorithms called quasi-Newton methods. In this particular case, a C port (Okazaki, 2002–2010) of Nocedal's reference Fortran implementation (Zhu et al., 1999–2014) was used.

4 Experiments

In this section, details of the hand configuration recognition system are given, along with details of the experiments, conducted to measure the performance of different aspects of the system. This also includes a detailed description of the footage used in these experiments.

The rest of this section is organized as follows: A very brief overview of the system and the three different stages of experiments is given in Section 4.1. Some implementation details are presented in Section 4.2. The footage used in the experiments is described in Section 4.3. Hand fitting and evaluation stages of the experiments are described in Sections 4.4 and 4.5, respectively. Section 4.6 describes how the distance data was used to form simple classifiers and how the classifier performance was measured.

4.1 Overview

The experiments were run in three stages. These stages are presented graphically in Figures 18–20. In the first stage, preprocessing steps were performed. These include skin detection and hand location detection, a simple application of methods from Sections 3.2–3.3. The stage is presented graphically in Figure 18.

In the next stage, the hand fitting stage, based on the hand location information, hand synthesis was performed, and the best pose of a given hand configuration, with respect to a given input file and a certain feature/metric/optimization combination, was sought. This was repeated for every input file, for a number of hand configurations, and for several different feature/metric/optimization combinations. At the end of each experiment run, the rendered image of the hand in its best pose was stored for later use, along with details of the best pose. This stage is presented graphically in Figure 19.

In the hand evaluation stage, the goodness of each best rendering was evaluated based on multiple features and metrics. At this point, the hand was not synthesized any more; the stored best rendering from the previous stage was used as input. The evaluation was done for every output from the hand fitting stage, and the results were stored for further analysis. Analysis of these results is presented in Section 5. This stage is presented graphically in Figure 20.

4.2 Implementation details

The experiments were carried out with the `slmotion` toolkit (Karppa et al., 2014). `slmotion` is a computer vision toolkit specifically designed for the automatic analysis and annotation of sign-language corpora. The program has been written in C++11 (ISO, 2011), and is built around the OpenCV library (Willow Garage, 1999–2014).

The toolkit provides users a framework for writing analysis components, and a pre-existing library of such components. The software allows its users to conduct experiments by exposing an Application programming interface (API) for the Python

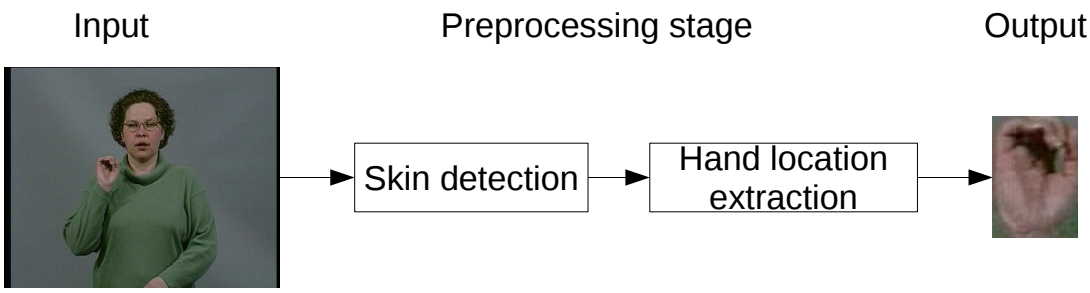


Figure 18: The preprocessing stage of the system. The hand location is extracted on the basis of skin detection results.

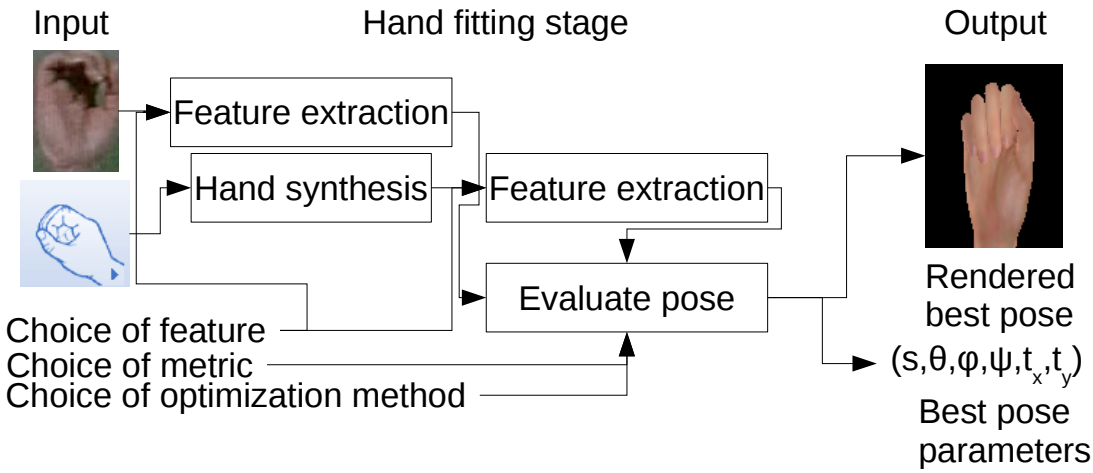


Figure 19: The hand fitting stage. The synthetic model of a given hand configuration is fitted to the input image, on the basis of a given feature/metric/optimization combination. The result is a rendered image of the configuration in the best pose, and the corresponding pose parameters.

programming language (Python Software Foundation, 1990–2014).

The toolkit has been used successfully for analyzing sign-language video material in the past. Viitaniemi et al. (2014) presented S-pot, a benchmark database for sign spotting along with a baseline solution to the sign spotting task. The baseline solution was implemented within the `slmotion` framework and the experiments were carried out with the toolkit. Before that, a method for detecting hand-head occlusions was implemented within the `slmotion` framework, along with experiments (Viitaniemi et al., 2013). Karppa et al. (2011) presented a method implemented in an early version of the toolkit for tracking hand and head movements. The performance of this method was later evaluated by applying the method to video data recorded in a setting where motion capture equipment was used for providing reference (Karppa et al., 2012). Data produced with this method has also been used for analyzing head movements in the Master's Thesis of Puupponen (2012).

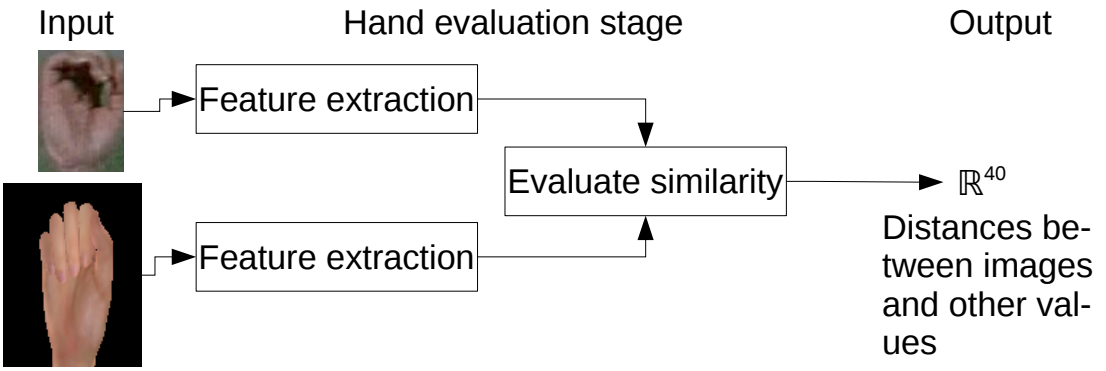


Figure 20: The hand evaluation stage. Several, possibly computationally very expensive, features are extracted from both the input and the rendered image of the synthetic hand. Several metrics are used to compute the distances between these features. The distances are then output as a feature vector for further analysis.

While `slmotion` itself parallelizes poorly, it is still possible to use it to analyze massive corpora, as in a typical case each file is analyzed separately. For instance, in the case of this work, the results were evaluated by running tens of thousands of instances of `slmotion` on a Condor[3] grid of some 400 nodes. Unfortunately, it was not possible to do the same for the hand fitting part which required the synthesis of hand images, as it turned out that the LibHand (Šarić, 2011) library that was used for the task required access to the X Window System, in order to exploit the GLX extensions. The cluster did not permit remote access to workstation X servers, so these runs had to be performed manually on a handful of workstations.

4.3 Test footage

Experiments were carried out using a select subset of frames from videos from Suvi, the on-line video dictionary of Finnish Sign Language (Finnish Association of the Deaf, 2003–2014). The particular version used was an older snapshot of original videos, provided directly by the Finnish Association of the Deaf. The very same snapshot of 1220 articles has been used as basis for the S-pot benchmark (Viitaniemi et al., 2014).

The Suvi dictionary consists of a number of dictionary articles, each describing a particular sign. Almost every sign contains a *citation form* video where the sign is shown in isolation. Also, each article contains one to four example sentences which show how the sign is used in a context. The articles are annotated with linguistic features that describe some of the properties of the sign. These include:

- the handshape (or hand configuration)
- place of articulation (POA)

[3] `http://research.cs.wisc.edu/htcondor/`

- number of active hands
- the mouthing
- type of movement (such as spinning motion, change of handshape, absence of motion, etc.)

The Suvi hand configuration classification was chosen for describing the abstract hand configurations for the hand fitting and evaluation task. This is because, to the author's best knowledge, no other equally exhaustive hand configuration classification exists for Finnish Sign Language. The author also believes that his linguistic expertise is not at a level high enough to disagree with the authors of the dictionary.

Suvi distinguishes between 36 different hand configuration categories. The articles are annotated with these classes. Authors of the dictionary also recognize the well-known fact that the hand configurations may manifest in many modified ways, but these minor handshapes have not been annotated in the articles. Hence, in all experiments, the suggested main hand configuration was adopted, lacking better knowledge.

The hand configurations are identified by four-digit codes. The codes have been adopted directly from Suvi. The first digit is always one. The second digit identifies a group of similar hand configuration classes; there are six groups, each consisting of 3–9 hand configurations. The groups are:

- Palm configurations (10xx, 6 in total)
- Fist configurations (11xx, 5 in total)
- One-fingered configurations (12xx, 3 in total)
- Two-fingered configurations (13xx, 9 in total)
- 3–5-fingered configurations (15xx, 7 in total)
- Grip configurations (16xx, 6 in total).

The third digit identifies the class in question. The last digit is zero for hand configuration classes with only one known manifestation, and one for the main manifestation of a hand configuration with multiple modified forms.

Out of the 36 different hand configuration classes, 26 were present in the frames chosen for experiments. Symbols for these hand configuration classes, extracted from Suvi, along with the synthetic representations, are shown in Appendix A.

It should be noted that the handshapes are not evenly distributed. Figure 21 shows the distribution. The most common handshape, 1001, occurs a total of 240 times, while certain handshapes, namely 1141, 1530, 1560, 1640, and 1650, only occur once.

Although the videos used were originals, and thus in better quality than the ones available on the Web, image quality was modest at best. Some of the oldest videos had been shot in the 1990's with analog Betacam cameras, and have since then been

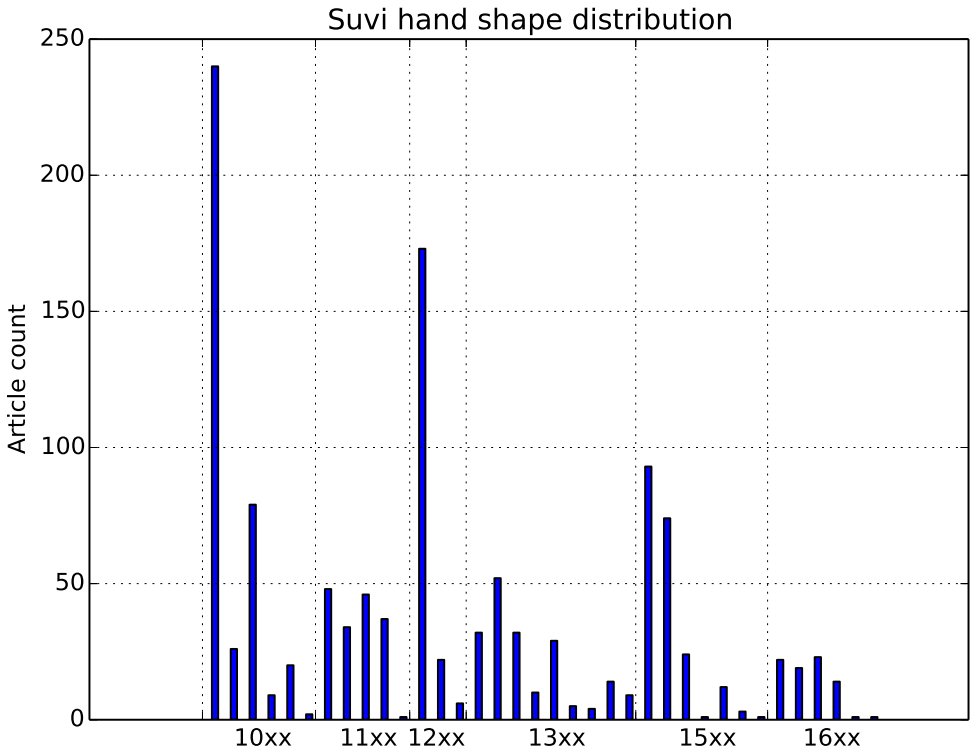


Figure 21: The distribution of handshapes among the 1220 Suvi articles.

digitized. The files provided by the Finnish Association of the Deaf were stored in Digital Video (DV) format. The DV compression (IEC, 1998) reduces the image quality even further. The video was interlaced, so the video was deinterlaced using the FFMPEG deinterlace method, which involves convolving one of the two fields vertically with a $\begin{bmatrix} -1 & 4 & 2 & 4 & -1 \end{bmatrix}$ kernel.

Owing to the deinterlace process, and the fact that the camera only recorded 25 frames per second, it is utterly infeasible to identify the handshape from most of the frames, even by a human observer. This is because hand movements tend to be very fast in comparison to the amount of time recorded per each frame. The image resolution is also poor, nominally 576 lines, but in practice lower because of the fact that the footage was originally recorded using an analog camera, and the fact that YUV 4:2:0 compression is used with the European Phase Alternating Line (PAL) DV format. In that format, color information is stored only at half the nominal resolution, ie. for a block of 4 pixels, 4 bytes are used for luminance information and 2 bytes for color information.

All experiments were conducted on a frame-by-frame basis, and no motion information was used. Therefore, the frames that were used in the experiments were selected very carefully, according to multiple criteria. The frames were chosen only among the citation form videos, as it is known that the hand configuration corre-

sponding to the article should appear in that video. In case of example sentence videos, it is not clear to a person who does not understand Finnish Sign Language, where exactly the sign begins and ends, and whether it is possible to accurately identify the handshape. While it would certainly have been feasible to also use frames from example sentences, this would have required tedious work, and was ruled to be beyond the scope of this work.

The second most important criterion was the sleeve length. In a majority of the videos, the person signing the citation form is wearing a short-sleeved shirt. While most of the techniques explored in this work are not affected by this, the hand location detector is. Geometric constraints on skin-colored regions do not provide a sufficiently accurate estimate of the location of the *hand*. As noted in Section 3.3, there would have to be a method for cutting the hand at the wrist, and since this is highly non-trivial, short-sleeved videos were simply excluded from the experiments. Short-sleeved and long-sleeved signers were easy to identify automatically, by applying skin detection and comparing the amount of skin-colored pixels in the bottom half of the first frame of the video. This in part was due to the fact the signers had been instructed to keep their hands in a neutral position, hanging towards the floor, at the start of the video, which meant that very little skin was visible, unless the signer was wearing a short-sleeved shirt.

Other criteria for exclusion included hand-head and hand-hand occlusions. The reasons are again related to hand location detection, as skin color detection is not a sufficient tool to disambiguate between the different overlapping regions. Furthermore, a great fraction of the frames were unusable because of a high level of motion blur. In the end, good frames were identified manually by going through the remaining videos frame-by-frame.

In the end, 237 frames from 166 videos and five different signers were selected. Figures 22a–22e show an accepted frame from each signer. The signer in Figure 22d is left-handed, so images of him were manually adjusted by mirroring them. Figures 22f–22i show rejected frames. The frames in Figures 22f and 22g were rejected because two skin regions are overlapping, causing the hand detection algorithm to fail. The frame in Figure 22h was rejected because of a high level of motion blur was present. The frame in Figure 22i shows a typical example of a video where the signer was wearing a short-sleeved shirt, the reason of rejection.

4.4 Hand fitting stage

After the hand location has been determined for each frame in the preprocessing stage, every hand configuration is fitted against every frame, using selected feature/metric/optimization combinations for determining the best pose. This is done to fix the remaining six degrees of freedom. The process is shown in Figure 19.

Details of how a single instance is fitted were presented in Section 3.11. Briefly, after the choice of the feature, the metric, and the optimization scheme is made, the feature is computed for the input frame. Then, the hand is rendered at different orientations, at given steps of angles. If no optimization is done, the corresponding features are simply computed and the metric is applied to compare the features.

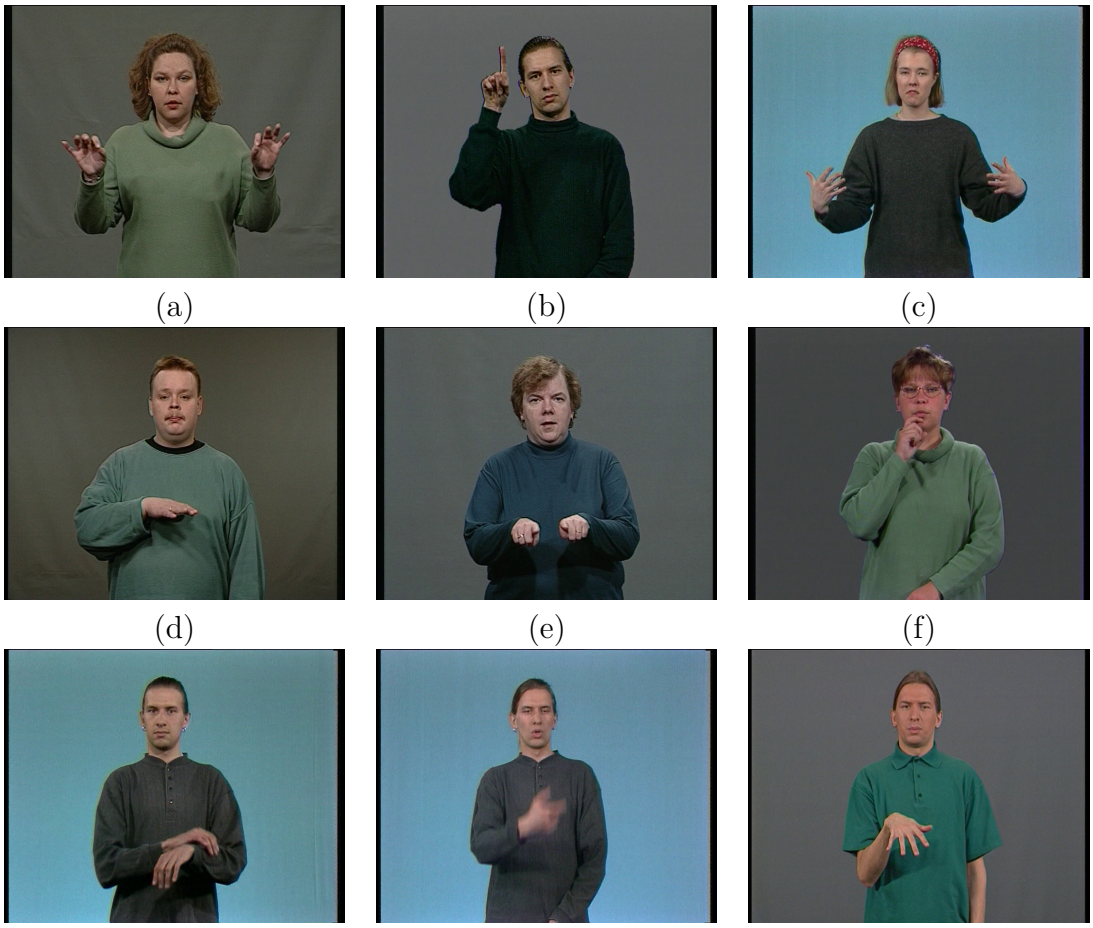


Figure 22: Frames extracted from the Suvi corpus. (a)–(e) are accepted. (f) and (g) are rejected because of overlapping skin regions. (h) is rejected because of motion blur. (i) is rejected because of short sleeves.

Otherwise, the angles are used to determine the initial configuration for the optimization scheme, the feature and metric combination is used to construct the cost function, and the optimization scheme is used to find the final configuration. Once this has been repeated for all possible angles, at the given step size, the configuration that minimized the value of the cost function is chosen, and the rendered image of the configuration at the given parameter values is passed to the next stage.

The two optimization schemes, L-BFGS and gradient descent, were described in Section 3.11. As the optimization schemes increase the runtime by an order of magnitude, if not more, the initial pose was only probed with a step size of 45°. The experiments without any optimization were repeated at step sizes of 45°, 30°, and 22.5°.

The feature/metric combinations used are shown in Table 2. They are called "scan" features and metrics because they are used for scanning the pose space for the best pose to evaluate the configuration. A key criterion for choosing these pairs

Table 2: Features and metrics used as "scan" features and metrics in the hand fitting stage of experiments. With the trimmed HOG and EMD, the HOG Bin Distance (see Section 3.7) was used as the ground metric.

Scan feature	Scan metrics
Outer contour	Chamfer
Canny edges	Chamfer
HOG	χ^2, Euclidean
PHOG (3 and 4 levels)	χ^2, Euclidean
PHOG (5 levels)	χ^2
Trimmed HOG (max size 10/50)	EMD, $\widehat{\text{EMD}}$
Trimmed HOG (max size 100)	EMD

Table 3: Actual trials that were run. The #HC field lists the number of hand configurations that were evaluated (a subset of the 26 configurations in total), and #Imgs the number of frames that were associated with these hand configurations (out of the 237 in total).

Feature/Metric	#HC	#Imgs
Contour/Chamfer	12	155
Canny/Chamfer	12	155
HOG/Euclidean	12	155
HOG/χ^2	2	68
PHOG (3/4/5 levels) / χ^2	2	68
PHOG (3/4 levels) / Euclidean	2	68
Trimmed HOG (size 10/50) / EMD	2	68
Trimmed HOG (size 100) / EMD	12	155
Trimmed HOG (size 10) / $\widehat{\text{EMD}}$	2	68
Trimmed HOG (size 50) / $\widehat{\text{EMD}}$	12	155

is that both the features and the distances must be computable very rapidly. This is because they need to be evaluated possibly for thousands of times during each trial run. This makes it impossible to apply the EMD to the vanilla-HOG, as the typical runtime would be in the order of hundreds of seconds. Therefore, the trimmed HOG was used with the EMD at various maximum sizes. Furthermore, the PHOG was tested at 3, 4, and 5 levels. Altogether, this yields 14 scan feature/metric pairs.

With five different pose space probing schemes, 14 scan feature/metric pairs, 26 hand configurations and 237 input frames, a total of 431,340 different trials were to be run. Due to time constraints, it was not possible to perform all of these trials in a reasonable amount of time. Another source of complication was that the LibHand library, used for rendering the synthetic hand images, required access to the X Window System. This meant that the experiments could not be run on a computational cluster, so the trials were run on only two workstations, with

Table 4: The hand configurations chosen for further tests. Max accuracy is the best discriminatory accuracy obtained in a binary classification test with manually assigned poses, using a nearest neighbor classifier. Avg. accuracy is the average accuracy over all nearest-neighbor classifiers. The hand configurations listed here contain both some of the easiest cases and the most difficult cases, in terms of discriminatory performance.

Hand configuration pair	# Imgs	Max accuracy	Avg. accuracy
1001/1201	42/29	88.7 %	68.0 %
1011/1031	3/2	100 %	78.3 %
1301/1521	9/2	100 %	75.0 %
1311/1341	11/5	100 %	70.8 %
1001/1130	42/21	98.4 %	73.4 %
1511/1521	20/2	54.5 %	31.4 %
1341/1511	5/20	56.0 %	32.3 %
1110/1601	9/7	56.3 %	37.5 %

eight Central Processing Unit (CPU) cores each. A summary of the trials that were actually run is presented in Table 3. The hand configurations were chosen on the basis of running pairwise binary classification experiments on each pair with manual poses, and then selecting hand configuration pairs which were manifested in non-trivial numbers, and which were easily discriminated, and, on the contrary, a choice of very poorly discriminated cases. The hand configurations selected are listed in Table 4. The scan feature/metric combinations were chosen by first running all possible combinations on the two most common hand configurations, and then selecting those that lead to good performance in the the binary classification case. The rendered best poses, best pose parameters, and optimization times were stored on disk for further analysis in the hand evaluation stage.

The three angular parameters resulting from this stage, θ, ϕ, ψ, were compared against manually annotated angles, which formed a baseline for the evaluations in the hand evaluation stage. These annotations were created by having a human observer go through every one of the 237 images, and form his or her own opinion on the pose of the hand, by having the observer synthesize the hand images manually. It should be noted that such annotations are inherently imprecise, owing to the fact that it is impossible to obtain the 3D pose from a 2D projection, as one degree of freedom is inevitably lost.

The angular differences were evaluated against the manually assigned angles in terms of Mean Square Error (MSE) where the error function $\mathcal{E}$ was defined as

$$\mathcal{E} = \sum_{\alpha \in \{\theta, \phi\, \psi\}} \left(\min\left(|\alpha - \alpha'|, 2\pi - |\alpha - \alpha'|\right)\right)^2 \tag{28}$$

where α, α' stand for the fact that one angle is taken from the output of the algorithm, and the other is the manually annotated equivalent.

An alternative characterization with values possibly easier to interpret was obtained by considering the angular values to constitute vectors of the $\mathbb{R}^3$ space, and

Table 5: Features and metrics used for evaluation. With the trimmed HOG and EMD, the HOG Bin Distance (see Section 3.7) was used as the ground metric. For SURF and SIFT descriptors and EMD, the Euclidean distance was used as the ground metric.

Feature	Metrics
Outer contour	Chamfer
Canny edges	Chamfer
HOG	χ^2, Euclidean
PHOG (5 levels)	χ^2, Euclidean
Trimmed HOG (up to 99 % of total mass)	EMD, $\widehat{\text{EMD}}$
SIFT	EMD, $\widehat{\text{EMD}}$
SURF	EMD, $\widehat{\text{EMD}}$

considering the angle between the two vectors. That is, letting $\mathbf{u} = (\theta, \phi, \psi)$ and $\mathbf{v} = (\theta', \phi', \psi')$ be the vectors, one the output of the algorithm and the other the ground truth, define

$$\xi = \arccos\left(\frac{\mathbf{u} \cdot \mathbf{v}}{||\mathbf{u}||_2 ||\mathbf{v}||_2}\right). \tag{29}$$

4.5 Hand evaluation stage

After the best pose for each of the hand configurations with respect to each input frame was determined in the hand fitting stage, the goodness of each fit was determined by computing a set of features for the input frame and the best rendered image of the configuration. The similarity of the images was evaluated by comparing these features by the means of different metrics. This process is shown in Figure 20. It should be noted that no hand synthesis was performed at this point.

The feature/metric combinations used in this stage are shown in Table 5. These include most of the features used for scanning the pose space, but also more expensive combinations, and more parameter values, such as an increased number of PHOG levels. It is possible to evaluate this increased number of feature and metric combinations because the evaluation needs only be carried out once.

As a result, a vector of 40 floating-point values was computed. The vector consists of

- the values and run times of the distance functions at each feature/metric combination in Table 5 ($2 \times 12 = 24$ values)
- the best pose parameters from the hand fitting stage (6 values)
- Canny Edge and SIFT/SURF interest point counts for the input image and the best rendered image (6 values)
- HOG, PHOG, and Trimmed HOG descriptor sizes for the hand and the rendered image (4 values)

The optimization times from the hand fitting stage were treated separately. In addition, the best rendered images were stored on the disk. As no hand synthesis was required at this point, the evaluation was performed on a HTCondor cluster with some 400 CPUs available. Despite the increase in computational resources, the evaluation took a considerable amount of time, as each trial would take up to 30 minutes to complete.

A baseline was computed by using the manual fits, as described in the previous section. This gives an idea of an upper limit that the system can achieve with a supposedly perfect hand fitting stage.

4.6 Classification experiment

The output of the hand evaluation stage is a vector consisting of a number of distances from synthetic hand configurations to the input images and other parameters and measured values. A simple way to construct a classifier in this manner is by the means of nearest neighbor search. In that scheme, each input video frame is associated with a hand configuration by choosing the one that minimizes one of the distances. As the ground truth is available from Suvi annotations, a simple way to evaluate the classifier would then be by simply counting the fraction of correctly classified input files.

Let $\mathcal{H}$ be the set of hand configurations, $\mathcal{F}$ the set of features, $\mathcal{M}$ the set of metrics, and $\mathcal{O}$ the set of fitting schemes (step size and optimization method). Let I be the input image. Let $d_{f,o}(h, I)$ be the distance from some hand configuration $h \in \mathcal{H}$ to the input image I, computed using metric $d \in \mathcal{M}$ with respect to feature $f \in \mathcal{F}$, using the fitting scheme $o \in \mathcal{O}$, ie. one component of the resulting vector of the hand evaluation stage. Then the nearest neighbor classification scheme would be

$$c_{NN,f,d,o}(I) = \arg\min_{h \in \mathcal{H}} d_{f,o}(h, I). \tag{30}$$

Another approach would be voting. In that case, the classification is performed using all features and metrics, and the hand configuration which minimizes most distances is chosen. That is,

$$c_{v,o}(I) = \arg\max_{h \in \mathcal{H}} \sum_{f \in \mathcal{F}, d \in \mathcal{M}} [c_{NN,f,d,o}(I) = h] \tag{31}$$

where $[\cdot]$ is taken to evaluate to 1 if the expression is true, and 0 otherwise.

For a classifier to have any credibility at all, it should be able to beat the dummy classifier. In this case, the dummy classifier would assign the most probable label, "1001", to every image. In essence, the dummy classifier can be thought to be a "maximum a priori" classifier.

In addition to simple accuracy figure as the fraction of correctly classified cases out of all cases, one way to measure the classifier accuracy is by considering binary classification. In that case, the classifier attempts to decide a simple YES/NO question of whether the given handshape corresponds to the proposed hand configuration. This allows one to consider precision and recall figures. These figures were

already defined in Equations (3) and (4). Precision measures the fraction of actual detections, while recall measures the proportion of detections out of all instances.

To determine whether classification errors arise from confusion between similar hand configurations, or whether they are completely arbitrary, the classification was also performed between hand configuration *groups*. The groups here correspond to the six different hand configuration groups, as described in Section 4.3 and visualized in Appendix A. A classification result is considered correct with respect to the hand configuration group if the predicted class belongs to the same group as the ground truth. This allows the groupwise classification to be treated as an information retrieval task, and, as such, the precision/recall/accuracy figures from Equations (3)–(5) become meaningful.

While there is nothing spectacular about the precision and recall figures, however, it should be noted that the accuracy values computed this way may look unreasonably favorable, as they are dominated by the vast number of true negatives. As such, to give a better understanding of the overall performance, let us construct a figure called *total groupwise accuracy*.

Groupwise accuracy is defined as the proportion of images, from a given group, correctly classified to the group. Formally, let $g \subseteq \mathcal{H}$ be a group of hand configurations, and $\mathcal{I}_g$ the set of images which depict a hand configuration from group g. Then,

$$a_{g,f,d,o} = \sum_{I \in \mathcal{I}_g} \frac{[c_{NN,f,d,o}(I) \in g]}{|\mathcal{I}_g|} \tag{32}$$

is the groupwise accuracy of group g with respect to feature f, distance function d, and optimization scheme o. Here $[\cdot]$ is again taken to evaluate to 1 if the expression is true, and 0 otherwise.

To incorporate a priori distribution information, let the *total groupwise accuracy* $\hat{a}_{f,d,o}$ be the fraction of images correctly classified with respect to their hand configuration group. Formally, let $\mathcal{G}$ be the set of all six groups. Then,

$$\hat{a}_{f,d,o} = \frac{\sum_{g \in \mathcal{G}} \sum_{I \in \mathcal{I}_g} [c_{NN,f,d,o}(I) \in g]}{\left| \bigcup_{g \in \mathcal{G}} \mathcal{I}_g \right|}. \tag{33}$$

5 Results

Results of the experiments are presented in this section. Due to their large size, some of the tables have been omitted from this section and presented as separate Appendices. The rest of this section is organized as follows: Accuracy of the fitted handshape poses in terms of MSE and the angle between vectors compared to the manually annotated principal axes of orientation is discussed in Section 5.1. Classification results with the full set of 26 hand configuration classes and a restricted set of 12 classes are presented in Sections 5.2 and 5.3, respectively. Binary classification results are shown in Section 5.4. Runtimes of the different parts of the system are considered in Section 5.5.

5.1 Accuracy of fitted pose angles

MSEs of fitted hand configurations, compared to the manually annotated ones, for all configurations, along with the average angles between angular vectors as defined in Equation (29), are shown in Table 6. The values have only been computed for the hand configurations that correspond to the ground truth. For comparison, similar values for the most common hand configuration alone, 1001, are shown in Table 7.

Considering that the error value could theoretically range anywhere from 0 to 29.7, the values show reasonable average agreement. However, this still leaves the possibility that the angles may be off, on average, 85°–122°, which does sound like a lot of variation. This can be verified from Figure 23 which shows the results of all fitting operations to the frame shown in Figure 24. There are some obviously good results, but also a lot of spurious results which differ greatly in some angles, possibly by as much as 180°. In the case of some of the simpler features, such as contours, this is to be expected. As even a small number of such errors can greatly increase the MSE, this may make the MSE too coarse an estimate of the level of error, and its distribution in particular. The lack of predictable, systematic differences in the MSE between the different optimization schemes suggests that this may be the case.

The coarseness of the MSE is verified by the histograms in Figures 25a–25d. The figures show the squared error histograms for the fitting schemes that minimize (Figures 25a and 25b), and maximize the MSE (Figures 25c and 25d). It can be seen that the distribution is very uneven. This may make the MSE look worse for the good schemes than can be justified. On the other hand, in the worst case, there are occasions where the values may be very close to the theoretical upper bound which dominate the MSE. This is very pronounced with the histogram of the contour-based fitting, where the hand is 180° degrees off in some cases, although from the point of view of the feature, the silhouette remains almost the same. Later, it will be shown that the contour-based features worked, indeed, quite well for discriminatory purposes, so this suggests that the actual, precise estimation of the pose angles and the classification of handshapes may be somewhat unrelated tasks.

Figure 23: The hand configuration 1001 fitted to image of Figure 24, using all scan feature/metric/optimization combinations.

Table 6: Average MSEs of hand pose angles, fitted with various methods, computed with respect to the manual annotations (in radians squared), and corresponding average angles (in degrees), as defined in Equations (28) and (29), respectively.

Feature/Metric	None/45°	None/30°	None/22.5°	GD	L-BFGS
Contour/Chamfer	10.45/40.1°	10.39/38.6°	10.91/40.3°	11.13/43.0°	9.01/33.5°
Canny/Chamfer	9.99/40.1°	10.03/38.0°	10.43/38.6°	10.98/43.5°	9.53/41.9°
HOG/χ^2	8.89/42.1°	9.68/35.9°	8.37/35.8°	10.33/36.6°	10.94/43.7°
HOG/Euclidean	8.84/38.5°	9.96/37.9°	8.87/36.3°	10.76/39.9°	8.64/43.3°
PHOG3/χ^2	9.63/35.2°	10.36/40.0°	10.41/39.3°	9.95/39.7°	10.35/38.9°
PHOG3/Euclidean	10.36/38.5°	8.69/41.9°	9.99/43.5°	9.43/47.2°	10.35/36.8°
PHOG4/χ^2	10.13/38.0°	8.73/36.9°	11.08/37.2°	9.07/39.0°	10.34/36.8°
PHOG4/Euclidean	11.07/36.5°	9.69/42.5°	11.35/38.6°	9.19/37.9°	10.82/35.3°
PHOG5/χ^2	9.06/36.5°	10.05/41.0°	9.68/36.9°	8.90/36.2°	10.52/40.8°
THOG/10/EMD	9.95/42.1°	9.48/41.4°	9.03/40.5°	10.62/47.5°	8.84/39.5°
THOG/10/$\widehat{\text{EMD}}$	9.75/42.2°	9.30/43.9°	9.86/42.1°	8.75/41.3°	8.51/39.6°
THOG/50/EMD	9.87/42.7°	9.40/42.9°	8.91/36.4°	10.40/41.1°	8.75/40.7°
THOG/50/$\widehat{\text{EMD}}$	8.74/37.2°	9.52/40.6°	9.72/39.1°	11.29/44.6°	8.91/36.7°
THOG/100/EMD	9.65/40.5°	10.08/39.2°	9.35/39.2°	10.61/44.7°	9.28/36.3°

Table 7: MSEs of hand pose angles for the most common hand configuration, 1001, fitted with various methods, computed with respect to the manual annotations (in radians squared).

Feature/Metric	None/45°	None/30°	None/22.5°	GD	L-BFGS
Contour/Chamfer	8.91/39.4°	8.99/34.8°	10.81/38.6°	10.18/48.7°	7.69/32.4°
Canny/Chamfer	10.28/42.1°	9.64/38.3°	9.87/39.2°	11.28/45.0°	10.64/39.8°
HOG/χ^2	9.82/43.4°	11.04/39.2°	7.96/34.9°	11.01/37.9°	10.09/39.3°
HOG/Euclidean	10.30/43.1°	9.92/40.5°	7.76/34.4°	10.66/39.8°	7.77/36.3°
PHOG3/χ^2	8.46/33.7°	10.11/39.7°	10.66/39.0°	8.81/40.0°	9.46/36.9°
PHOG3/Euclidean	9.95/35.9°	7.10/39.3°	9.31/45.5°	8.46/45.4°	9.97/35.0°
PHOG4/χ^2	8.87/33.5°	8.36/38.3°	10.78/35.4°	8.70/39.2°	9.55/34.0°
PHOG4/Euclidean	9.89/31.9°	9.02/37.7°	11.35/39.9°	7.18/29.9°	9.55/31.0°
PHOG5/χ^2	8.66/33.3°	9.20/40.9°	9.17/36.2°	8.95/37.6°	10.36/39.7°
THOG/10/EMD	9.42/42.7°	8.93/41.2°	9.69/37.0°	11.66/48.0°	8.73/39.8°
THOG/10/$\widehat{\text{EMD}}$	9.25/39.6°	9.41/44.7°	10.38/39.3°	8.32/39.1°	8.85/38.1°
THOG/50/EMD	9.19/39.2°	9.08/39.4°	7.69/34.7°	11.62/38.5°	8.97/40.2°
THOG/50/$\widehat{\text{EMD}}$	8.93/37.3°	8.32/37.7°	8.35/36.6°	10.18/43.0°	9.08/35.0°
THOG/100/EMD	9.11/40.0°	9.28/37.5°	7.61/39.1°	10.66/47.5°	10.84/40.1°

5.2 Classification: Full set of 26 classes

With all 26 classes in play, the dummy classifier was formed by assigning each file the label "1001". This yielded a classifier whose accuracy turned out to be 17.7 %. Due to constraints described in Section 4.4, it was only possible to perform full classification task with manually adjusted pose. Considering the dummy classifier as a baseline, the results for the nearest neighbor classifier with manually fitted poses are shown in Appendix B.

In this case, the only feature/metric combination that was able to beat the

Figure 24: A frame from Suvi article 137, "TRANSLATE", showing the hand configuration 1001.

dummy classifier was PHOG/χ^2, and only by a small margin. This suggests that the full 26-class scenario may be too difficult for the simplistic approach. Also, the good performance of the dummy classifier is likely to be explained by the fact that the distribution of the handshapes in the data is very skewed. Simple majority voting, as per Equation (31), yielded an accuracy of 18.6 %, precisely equal to that of PHOG/χ^2, suggesting that the classifier dominated the voting in terms of correctness, with other classifiers only offering random support.

The situation does not look quite as grim if hand configuration groups are considered, rather than individual classes. The full nearest neighbor classifier accuracy matrix, for different groupwise classifiers, is shown in Appendix B. The matrix shows that, in terms of total groupwise accuracy, as defined in Equation (33), the dummy classifier performed at an accuracy of 28.7 %, and was beaten by contour/chamfer, Canny/chamfer, and Trimmed HOG/$\widehat{\text{EMD}}$ nearest-neighbor classifiers. This suggests that simple geometric cues, relating to the silhouette of the hand, were most effective. In terms of precision and recall, the dummy classifier performed very poorly, and was beaten by virtually every classifier.

There were striking differences between the groups. In addition to being the most common, the group of hand configuration classes 10xx, seemed to be the easiest, with multiple classifiers reaching over 50 % recall and precision. On the other hand, in the case of the group of hand configuration classes 16xx, there were classifiers which failed to classify any input images correctly. In terms of average precision and recall, the contour/chamfer, PHOG/Euclidean, PHOG/χ^2, and THOG/EMD classifiers performed reasonably well. The SIFT and SURF based classifiers performed rather poorly on all accounts.

Figures 26a–26e show confusion matrices of selected nearest-neighbor classifiers graphically. The values from which the images have been created are shown in Appendix C. The pixel values have been set logarithmically, with intensities ranging from 0 to 255, scaled such that if c_{ij} is the number of classifications of an instance

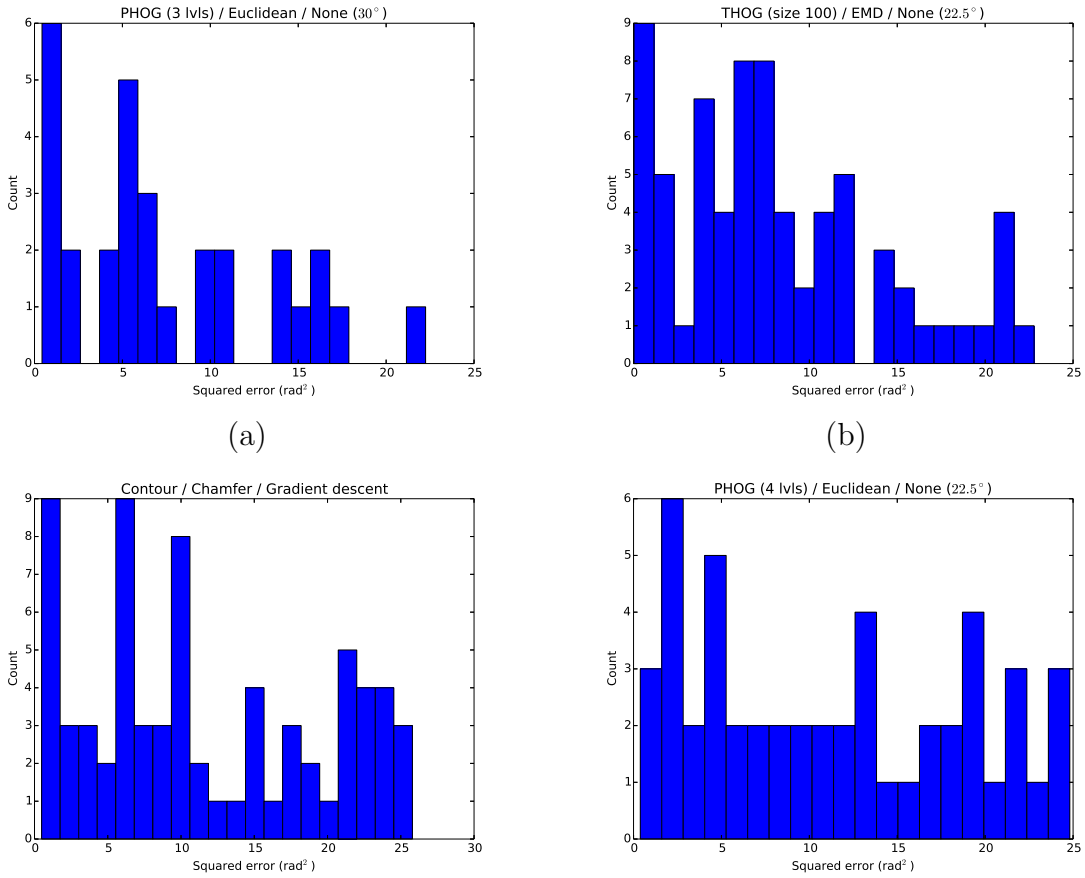


Figure 25: Histograms of squared angle errors, compared to the ground truth. (a) and (b) have minimal MSE, while (c) and (d) have maximal.

of class i (ground truth, vertical axis) classified as class j (prediction, horizontal axis), then the intensity value 255 corresponds to $\max_{\{c_{ij}\}} \log_2 c_{ij}$. The colors have been selected such that diagonal elements are green, corresponding to correct predictions, intra-group misclassifications, that is, incorrect classification but the predicted class being from the same group, are yellow, and inter-group misclassifications are red. Formally, let

$$v_{ij} = \frac{\log_2 c_{ij}}{\max\limits_{\{c_{k\ell}\}} \log_2 c_{k\ell}} \tag{34}$$

and if $g_i, g_j \in \mathcal{G} \subseteq \mathcal{P}(\mathcal{H})$ are the groups associated with classes i and j, then

$$I(i,j) = (R,G,B) = \begin{cases} (0, v_{ij}, 0) & \text{if } i = j \\ (v_{ij}, v_{ij}, 0) & \text{if } i \neq j \wedge g_i = g_j \\ (v_{ij}, 0, 0) & \text{otherwise.} \end{cases} \tag{35}$$

In each case, the distribution is rather spread out, and dominated by a large number of correct classifications of the most common hand configuration, 1001.

Table 8: Precision matrix between handshapes and features in the manual setting. Explanation of abbreviations: HC = Hand Configuration, Ct = Contour, Cn = Canny, Cf = Chamfer, Eu = Euclidean, PH = PHOG, TH = Trimmed HOG

HC	Ct/Cf	Cn/Cf	HOG/Eu	HOG/χ^2	PH/Eu	PH/χ^2	TH/EMD	TH/$\widehat{EMD}$
1001	0.52	0.56	0.95	0.95	0.81	0.87	0.54	0.59
1011	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00
1021	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1031	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1041	0.00	0.00	0.50	0.50	0.05	0.29	0.00	0.00
1100	0.18	0.00	0.07	0.09	0.00	1.00	0.25	0.03
1110	0.00	0.00	0.00	0.00	0.23	0.14	0.00	0.00
1120	0.00	0.67	0.17	0.20	0.67	0.67	0.00	0.10
1130	0.25	0.50	0.28	0.30	0.00	0.11	0.20	0.08
1201	0.57	1.00	0.50	0.00	0.60	0.88	0.50	0.00
1210	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00
1221	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1301	0.50	0.00	0.00	0.00	0.04	0.07	0.00	0.00
1311	0.25	0.20	0.00	0.33	0.11	0.00	0.00	0.09
1321	0.00	0.00	0.00	0.07	0.03	0.04	0.00	0.00
1331	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.00
1341	0.11	0.07	0.10	0.10	0.00	0.00	0.00	0.00
1371	0.20	0.00	0.00	0.00	0.00	0.00	0.12	0.00
1381	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1501	0.11	0.29	0.27	0.30	0.42	0.36	0.15	0.20
1511	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.60
1521	0.11	0.12	0.00	0.00	0.00	0.00	0.00	0.00
1601	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1610	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1621	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25
1631	0.00	0.00	0.00	0.00	0.06	0.05	0.00	0.00

There does not seem to be a clear, systematic pattern of error; the misclassifications seem to be spread out all over the class space, with low frequency.

Tables 8 and 9 show the precision and recall matrices, respectively, between each hand configuration and discriminatory feature (excluding SIFT and SURF features), computed with manually adjusted poses. The large number of zeros in both tables shows that the performance of the classifier has been very uneven. On the other hand, it also tells of the fact that the number of occurrences of most hand configurations was very low. Table 8 verifies the observation that the PHOG/χ^2 combo performed quite adequately. In particular, the precisions for the most common handshapes, 1001 and 1201, were very high, 0.87 and 0.88, respectively. This suggests that in cases where the instances are abundant, the classifier is highly reliable, making very few false detections. The problems are more with reliable identification, as can be seen in the recall matrix of Table 9.

Figure 27 shows the ROC curve of the PHOG/χ^2 classifier for the two most common hand configuration classes, 1001 and 1201. The ROC curve was computed by adjusting a distance threshold parameter from the minimum value exhibited to the

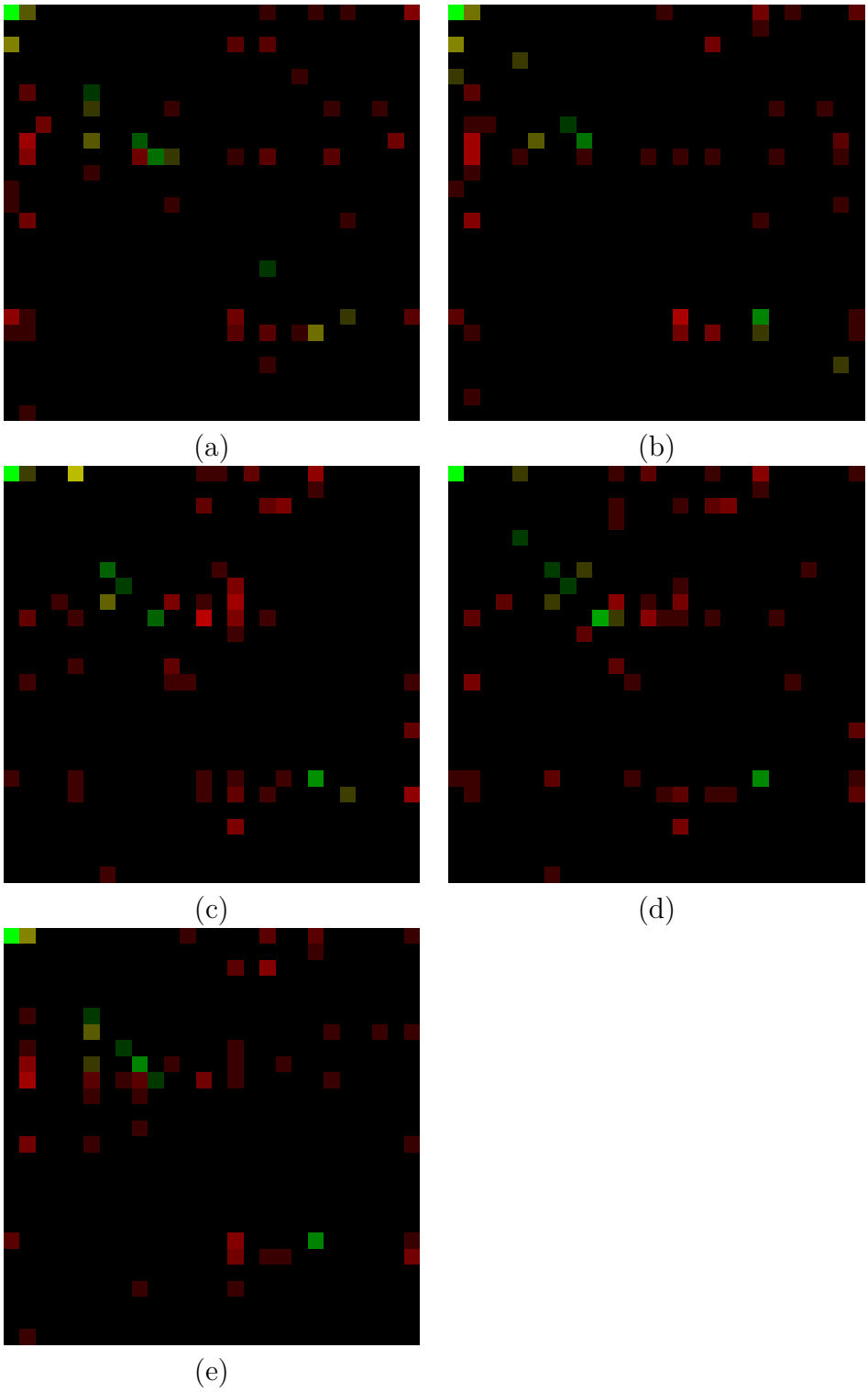


Figure 26: Confusion matrices of certain nearest neighbor classifiers with manually adjusted poses. (a) Contour/chamfer, (b) Canny/chamfer, (c) PHOG/Euclidean,, (d) PHOG/χ^2, (e) Majority vote. The numerical values from which these images were created are shown in Appendix C. The RGB pixel values have been assigned according to Equation (35).

Table 9: Recall matrix between handshapes and features in the manual setting. Explanation of abbreviations: HC = Hand Configuration, Ct = Contour, Cn = Canny, Cf = Chamfer, Eu = Euclidean, THOG = Trimmed HOG, PH = PHOG, TH = Trimmed HOG

HC	Ct/Cf	Cn/Cf	HOG/Eu	HOG/χ^2	PH/Eu	PH/χ^2	TH/EMD	TH/$\widehat{EMD}$
1001	0.57	0.52	0.45	0.48	0.40	0.48	0.50	0.24
1011	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00
1021	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1031	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1041	0.00	0.00	0.25	0.25	0.25	0.50	0.00	0.00
1100	0.33	0.00	0.33	0.33	0.00	0.17	0.17	0.50
1110	0.00	0.00	0.00	0.00	0.33	0.22	0.00	0.00
1120	0.00	0.20	0.10	0.10	0.20	0.20	0.00	0.10
1130	0.14	0.19	0.43	0.43	0.00	0.05	0.05	0.10
1201	0.14	0.03	0.03	0.00	0.10	0.24	0.07	0.00
1210	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00
1221	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1301	0.11	0.00	0.00	0.00	0.11	0.11	0.00	0.00
1311	0.09	0.09	0.00	0.09	0.09	0.00	0.00	0.09
1321	0.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00
1331	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
1341	0.40	0.20	0.20	0.20	0.00	0.00	0.00	0.00
1371	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
1381	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1501	0.05	0.23	0.14	0.14	0.23	0.23	0.14	0.05
1511	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15
1521	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00
1601	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1610	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1621	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50
1631	0.00	0.00	0.00	0.00	0.25	0.25	0.00	0.00

maximum, and evaluating the number of true and false positives and negatives, classifying the image to the respective class whenever the threshold was not exceeded. While definitely not ideal, the curvature is that of a working classifier.

5.3 Classification: A restricted set of 12 classes

A set of 12 hand configuration classes was selected for a binary classification task. The performance of the individual binary classifiers and the way the 12 classes were selected will be discussed in Section 5.4. However, a set of experiments similar to the full 26-class case were also run on the restricted set. The results of these experiments are shown in Appendix E, which contains nearest-neighbor matrices that are analogous to that of Appendix B.

As the set of input images was more restricted, hand poses fitted with other schemes could be considered in addition to manually adjusted pose angles. The different schemes are shown in separate tables. In the case of manually adjusted

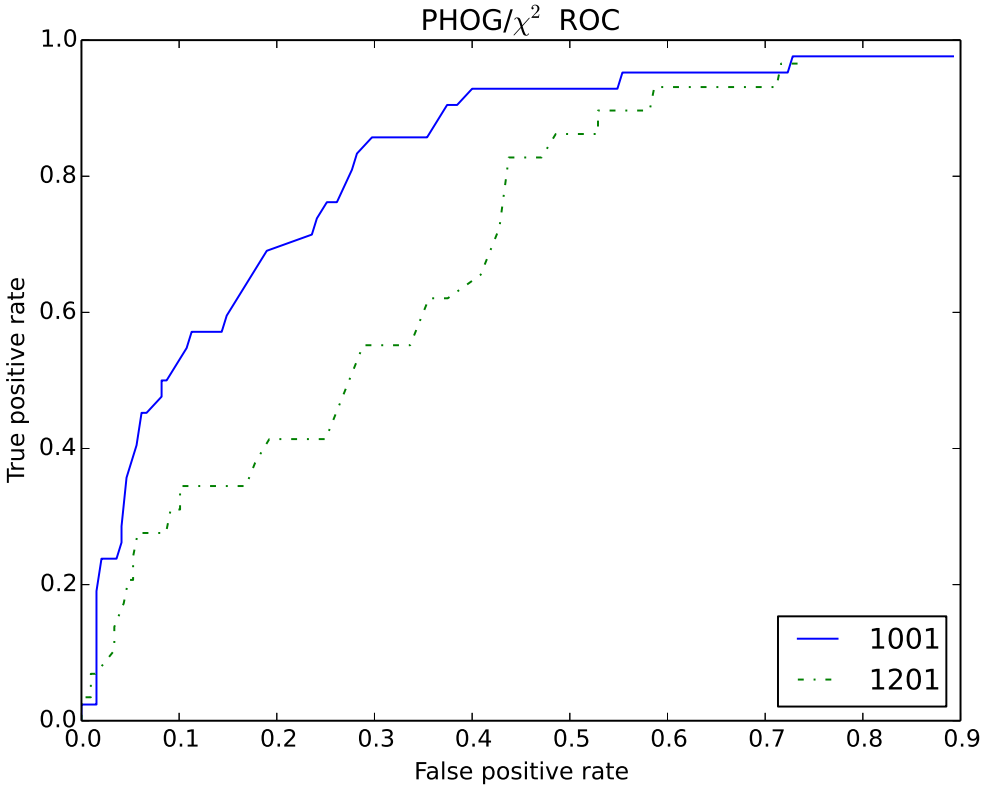


Figure 27: ROC curve for the PHOG/χ^2 nearest neighbor classifier, with manually assigned poses.

angles, overall, the performance was slightly better than in the 26-class case, which was to be expected. In the case where the degrees of freedom were fixed using the HOG features and the Euclidean distance, the expected pattern of improved classification accuracy as a function of decreasing step size arises. Furthermore, the gradient descent approach offers a great increase in classification accuracy, so much so that, while none of the classifiers reaches the 30 % level obtained with the contour/chamfer classifier in the manually adjusted case, the majority vote classifier in the HOG/Euclidean/Gradient descent case beats that of the manually adjusted poses by a margin of 0.6 percentage points, which can be viewed as a great success. Alas, despite the high runtime requirements, such improvements cannot be seen with L-BFGS optimization.

5.4 Binary classification

Another viewpoint to the classifier performance can be obtained by considering a binary classification case. The accuracy of the classifier with the manually set poses with the two most common classes 1001/1201, and a selection of very well and very poorly discriminated classes are shown in Table 10, with class pairs as columns

Table 10: Nearest neighbor classifier accuracies for all 26 classes, the two most common classes (1001/1201), some of the best discriminated classes with a non-trivial number of occurrences (1011/1031, 1301/1521, 1311/1341, 1001/1130), and some of the worst (1511/1521, 1341/1511, 1110/1601), using manually fitted hands. Abbreviations: Cf = Chamfer, THOG = Trimmed HOG, Eu = Euclidean.

Feature/Metric	all 26	1001/1201	1011/1031	1301/1521	1311/1341	1001/1130	1511/1521	1341/1511	1110/1601
Contour/Cf	16.9 %	88.7 %	80.0 %	100.0 %	75.0 %	98.4 %	13.6 %	24.0 %	43.8 %
Canny/Cf	16.0 %	80.3 %	80.0 %	100.0 %	93.8 %	90.5 %	13.6 %	20.0 %	37.5 %
HOG/Eu	15.6 %	73.2 %	100.0 %	81.8 %	100.0 %	76.2 %	27.3 %	24.0 %	31.3 %
HOG/χ^2	16.5 %	74.6 %	100.0 %	72.7 %	100.0 %	79.4 %	31.8 %	20.0 %	37.5 %
PHOG/Eu	15.2 %	73.2 %	100.0 %	81.8 %	81.3 %	85.7 %	27.3 %	32.0 %	25.0 %
PHOG/χ^2	18.6 %	81.7 %	100.0 %	81.8 %	81.3 %	92.1 %	22.7 %	28.0 %	25.0 %
THOG/EMD	13.1 %	83.1 %	100.0 %	81.8 %	68.8 %	92.1 %	9.1 %	20.0 %	31.3 %
THOG/$\widehat{\text{EMD}}$	9.7 %	62.0 %	80.0 %	63.6 %	50.0 %	57.1 %	50.0 %	28.0 %	37.5 %
SIFT/EMD	4.6 %	52.1 %	20.0 %	36.4 %	37.5 %	50.8 %	54.5 %	56.0 %	43.8 %
SURF/EMD	4.2 %	57.7 %	80.0 %	45.5 %	68.8 %	65.1 %	27.3 %	36.0 %	31.3 %
SIFT/$\widehat{\text{EMD}}$	4.6 %	52.1 %	40.0 %	63.6 %	37.5 %	60.3 %	54.5 %	48.0 %	56.3 %
SURF/$\widehat{\text{EMD}}$	5.9 %	36.6 %	60.0 %	90.9 %	56.3 %	33.3 %	45.5 %	52.0 %	50.0 %
Majority vote	18.6 %	71.8 %	100.0 %	81.8 %	81.3 %	85.7 %	27.3 %	24.0 %	31.3 %
Dummy	17.7 %	59.2 %	60.0 %	81.8 %	68.8 %	66.7 %	90.9 %	80.0 %	56.3 %

Table 11: Some of the best and worst discriminated classes, with their counts and maximum and mean classification accuracies with manually adjusted poses. The pair 1001/1201 is included because the two hand configurations are the most common.

	1001/1201	1011/1031	1301/1521	1311/1341	1001/1130	1511/1521	1341/1511	1110/1601
Max accuracy	88.7 %	100.0 %	100.0 %	100.0 %	98.4 %	54.5 %	56.0 %	56.3 %
Mean accuracy	68.3 %	80.0 %	75.5 %	71.6 %	74.4 %	31.1 %	31.7 %	37.0 %
Instance counts	42/29	3/2	9/2	11/5	42/21	20/2	5/20	9/7
Vote accuracy	71.8 %	100.0 %	81.8 %	81.3 %	85.7 %	27.3 %	24.0 %	31.3 %
Dummy accuracy	59.2 %	60.0 %	81.8 %	68.8 %	66.7 %	90.9 %	80.0 %	56.3 %

and different classifiers as rows. Results for the full 26 class case are included for reference. Some details of the class pairs are listed, along with their counts and maximum and mean classification accuracies, and voting accuracies, in Table 11. More binary classification results can be seen in Appendix F, which lists tables analogous to Table 10, for each scan feature/metric/optimization scheme applied to these 12 classes. In the appendix, the accuracies of the full 12-class classifier are included for reference.

The tables show that nearly perfect classification accuracies are achievable with good pose selection with some classes. In this reduced case, several of the classifiers beat the dummy classifier by a large margin. The PHOG/χ^2 performed very well. The pattern noted at the end of Section 5.2, of increased performance with the decrease of step size and, in particular, when the gradient descent optimization is used, is profoundly manifested in binary case as well. The HOG/Euclidean and HOG/χ^2 classifiers beat their equivalents in the case of manually adjusted poses, when the poses were set using the HOG/Euclidean/Gradient descent combination, by a margin of 4 percentage points, in the case of the pair 1001/1201. Majority vote

performance was equal. As before, the L-BFGS performance was rather poor.

A full accuracy matrix for the different fitting schemes with respect to different discriminatory features/metrics is shown in Appendix D, for the case of the most common pair 1001/1201. The results here were used for selecting the fitting schemes applied to the other full 12-class case. In particular, HOG/Euclidean, Canny/chamfer, Trimmed HOG (size 100)/EMD, and Trimmed HOG (size 50)/$\widehat{\text{EMD}}$ performed moderately well, with some discriminatory classifiers reaching accuracies in the excess of 70 %, particularly when gradient descent optimization was used.

5.5 Runtime analysis

Table 12 lists average runtimes of the hand pose selection stage with different feature/metric combinations used to form the objective function. As expected, the optimization schemes increase the runtime by an order of magnitude, or two. In particular, when comparing gradient descent, with the exception of PHOG5/χ^2, L-BFGS tends to take an order of magnitude longer than gradient descent. Given the poor results of L-BFGS optimization, this can hardly be justified.

The results also show that the runtime of the optimization schemes is highly sensitive to the time needed to evaluate the gradient of the objective function. Because of this, if the feature and the metric are simple to compute, simple probing with a smaller step size may be more expensive than either of the optimization schemes, and vice versa. The may occur if the gradient is computationally expensive, as in the case where multiple HOGs need to be evaluated. While not visible in the table, it should be noted that since the fitting was done on multiple computers and under a variable workload, there is a high level of variation in the runtimes, as the runtimes from different trials are not directly comparable. The averages should give insight into general runtime properties, however.

In a lot of cases, the runtimes are in the excess of thousands of seconds. For large corpus processing, this is a prohibitively large amount of time. Fortunately, some of the best performing feature/metric combinations, such as HOG/Euclidean and Canny/chamfer, are among the cheapest, which gives hope that, with a smarter probing of the pose space, the runtimes could be cut down to reasonable numbers.

Runtimes of the distance functions used in the evaluation stage are listed in Table 13. The table shows that the only distance functions that take a significant time to compute are the EMDs of HOGs.

Table 12: Runtimes of the best pose selection stage with different feature/metric combinations in seconds. Abbreviations: GD = "gradient descent", THOG n = Trimmed HOG with a maximum size of n.

Feature/Metric	None/45°	None/30°	None/22.5°	GD	L-BFGS
Contour/Chamfer	39.44	148.04	289.27	147.94	202.73
Canny/Chamfer	40.99	151.90	294.90	278.50	796.07
HOG/χ^2	38.71	145.76	284.44	244.46	5788.54
HOG/Euclidean	35.69	130.90	255.35	459.68	3670.67
PHOG3/χ^2	173.75	667.47	1268.62	6582.80	1920.03
PHOG3/Euclidean	192.30	707.65	1377.83	7439.53	759.25
PHOG4/χ^2	244.62	880.95	1678.66	9192.90	2702.56
PHOG4/Euclidean	231.81	866.21	1685.24	8823.29	984.91
PHOG5/χ^2	276.08	1058.27	2085.10	11081.07	2769.69
THOG 10/EMD	36.63	139.56	267.13	317.12	336.48
THOG 10/$\widehat{\text{EMD}}$	34.41	130.00	254.85	345.74	753.60
THOG 50/EMD	34.57	131.39	256.67	623.63	1984.65
THOG 50/$\widehat{\text{EMD}}$	40.77	149.94	290.56	675.42	7928.63
THOG 100/EMD	39.14	144.96	282.10	925.31	5753.82

Table 13: Runtimes of the distance functions used in the evaluation stage. Feature extraction times are not included.

Feature/Metric	Runtime (ms)
Contour/Chamfer	0.46
Canny/Chamfer	0.43
HOG/χ^2	0.40
HOG/Euclidean	0.11
PHOG/χ^2	0.04
PHOG/Euclidean	0.01
THOG/EMD	50124.65
THOG/$\widehat{\text{EMD}}$	48818.26
SIFT/EMD	0.48
SIFT/$\widehat{\text{EMD}}$	0.31
SURF/EMD	3.50
SURF/$\widehat{\text{EMD}}$	2.90

6 Discussion

While the results presented in the previous section may appear modest at first, it should be noted that the footage that was used to evaluate the system was very challenging. The fact that the approach presented here was able to classify handshapes correctly under favorable circumstances means that the system and the underlying linguistic model work to some extent. It is very likely that, given better footage and some of the improvements proposed later in this section, the system could be of practical significance.

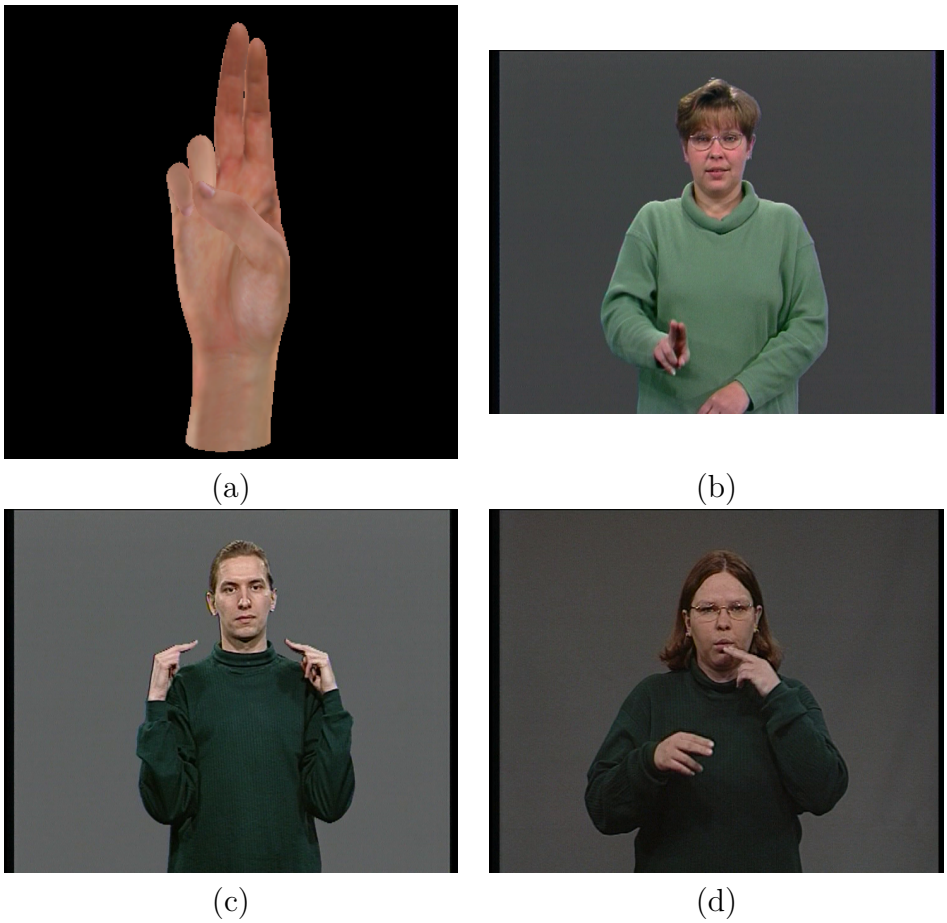
It was found that the HOG features and Canny edges can be used with the Euclidean and chamfer distances, respectively, to fix the degrees of freedom with a moderate success rate. This is a good piece of news as these features and distances are well-known, robust, and very simple to compute. The Trimmed HOG with the EMD and the HOG bin distance as ground metric also worked decently. However, despite an increase in runtime, the best simpler methods were not beaten. As the gradient descent optimization provided a great improvement in fixing the degrees of freedom, other, similar but computationally less demanding methods could make it possible to probe the pose space in a reasonable amount of time.

While the nearest-neighbor classifiers had, under favorable circumstances, high success rates, a lot of this is attributed to the fact that simple geometric cues, such as those provided by the contour, are often enough to discriminate between classes. The methods explored in this work do not seem to be adequate for addressing the cases where the silhouette cannot be used as aid, such as in the case of the fist handshapes. Altogether, it appears that, for a general-purpose system, multiple approaches need to be taken simultaneously for fixing the degrees of freedom of the hand model. This is because it is well possible that the methods that serve to disambiguate these difficult cases may not work quite so well with the simpler cases that can be reliably distinguished on the basis of their contour alone.

There are a number of ways that could possibly increase classification accuracy and decrease computational cost. Presently, pose space probing is done very inefficiently, and no use is made of the fact that the content is expected to be sign language. It is not an unreasonable expectation that the hand pose distribution is not uniform in the case of sign language. One possible way to extract this distribution could be by the means of cybergloves (Dilsizian et al., 2014), or motion capture equipment. University of Jyväskylä is known to have the equipment, and the necessary expertise (Jantunen, 2012; Karppa et al., 2012), but a large amount of data is required for estimating the a priori distribution reliably. The effort required for collecting the data is considerable.

Presently, the pose space was probed linearly. It would have been insightful to see if similar results would have been obtained with a randomized algorithm. Randomization could potentially have given good control over desired accuracy with respect to computational time, better than the step size parameter used in this work.

The behavior of the cost functions used also warrants further investigation. It would be interesting to see how many local minima there are, and how close they are to one another. This could give vision as to how the initial pose should be



(a) (b)

(c) (d)

Figure 28: (a) Rendering of hand configuration 1301. (b)–(d) Three different manifestations of hand configuration 1301.

selected. Also, the behavior of the gradient should be explored, and how it should be approximated. As the gradient is computed by rendering changes along all pose parameter axes, and taking the difference quotient, it is necessarily the case that too small changes cannot affect the image, as the rendering process is digital and thus quantized along all axes.

Classification results could also have been improved by taking in account the fact that the different hand configurations tend to be manifested in very different ways. For example, Figure 28a shows the rendered version of the hand configuration 1301, while Figures 28b–28d show three different manifestations of the same hand configuration. Although the handshape in Figure 28b is quite similar, the manifestations in Figures 28c and 28d differ wildly. The most striking difference is the bending of the index and middle fingers in their metacarpo-phalangeal (MCP) joint in Figure 28c. While these differences may seem meaningless to a human observer, they can have a dramatic effect when computing the distances. Using multiple hypothetical renderings per hand configuration could thus improve the results.

The method used for detecting the location of the hand is very simplistic. As the

later stages of the system are largely independent of the hand detection methodology, it is easy to incorporate more advanced hand tracking methods to the preprocessing stage of the system. This would permit the evaluation of the system in even more challenging conditions. The reported success of arm-tracking systems (Buehler et al., 2008, 2011; Pfister et al., 2012) suggests that larger context needs to be taken in account if the hand is to be located reliably even in the presence of occlusions. Information about the pose of the arm would enable the system to deduce where the wrist may be. This, in turn, would allow the system to rule out infeasible pose hypotheses, such as some of those in Figure 23 which have turned almost full 180 degrees. The search space for the correct pose would also be greatly reduced. This line of work warrants further investigation.

The first attempt at running the experiments was ruined by the fact that the bounding box was used as the size of the rendered image. This turned out to be too small. The resolution did not permit enough discriminatory power for the features, and the classification results were only little better than the toss of a coin. Furthermore, it turned out that LibHand renders such small and non-square images poorly, and there were numerous geometric anomalies in the output images. This was solved by simply adopting a 400×400 render window, and the input hand image was padded with zeros to fit this size.

Initially, the vanilla HOG was going to be used in evaluation with the EMD metric. Unfortunately, this turned out to be a bad idea as it turned out that a signed integer overflow bug in the OpenCV library prevented the use of their implementation of the EMD if the input vectors are very large. Furthermore, in high resolution experiments the HOG vector size would be on the order of 8×10^5 elements, which yields a floating-point distance matrix with a size in the order of 2^{33} bytes, or dozens of gigabytes, which is more than contemporary computers can handle. Fortunately, the trimmed HOG solved these problems.

In some cases, background clutter can cause problems with some features. One way to rectify this could be if the skin mask were detected robustly enough so that background can be removed. For this to work in the case of occlusions, a robust method needs to be available that can segment the hands from the skin area being occluded. Something along the lines of Viitaniemi et al. (2013) could be beneficial. As the test frames were chosen in such a manner that no occlusions were present, the non-skin pixels were simply eliminated. This improved classification results greatly.

While the lack of training can be considered a plus, it is evident that machine learning systems ought to be used to increase the classification performance, particularly in the last stage. Dilsizian et al. (2014) used Twin Gaussian Processes to map the appearance of the synthetic hand to the appearance of hands whose hand configuration class was known. Something along these lines could make the nearest-neighbor style classification more robust. The design of such systems needs to be done carefully, however, so as not to sacrifice signer independence.

The SIFT and SURF classifiers performed very poorly. This is likely due to the fact that the key point *location* was not taken into account in any way when computing the distances.

7 Conclusions

A computer vision system was presented that can, under favorable circumstances, locate the dominating hand from an individual sign-language video frame, and classify the hand shape using a synthetic 3D model. The system requires no training data; only an abstract, phonetically-motivated descriptions of the sought hand configuration classes are required.

Experiments were conducted to test various features and metrics to match the synthetic image to the input frame. In the first set of experiments, different features, metrics, and optimization schemes were used for scanning the 3D pose space for a likely pose of a given hand configuration, given the input frame, in terms of scale and the three angles of rigid body orientation in the 3D space. The goodness of the fits, in terms of recovering the raw pose of the three Euler angles, was compared to manually annotated ground truth. HOG features turned out to be suitable for this purpose with the Euclidean distance, and the Earth Mover's Distance, with a novel *HOG bin distance* as ground distance. Canny edges also performed favorably with the chamfer distance. Using gradient descent to minimize a cost function built from these measures improved results significantly.

In another set of experiments, the manually annotated poses were used as a baseline for building nearest-neighbor classifiers, using various feature and metric combinations, matching the rendered pose to the input frame. In a full scenario of 26 hand configuration classes, the system did not produce viable results, beating the dummy classifier only by a small margin with a PHOG/χ^2 classifier. On the other hand, it should be noted that the hand configuration distribution was very skewed, meaning that the dummy classifier worked particularly well.

The results were better when groups of phonetically similar hand configurations were concerned, with contour/chamfer, Canny/chamfer, and PHOG/χ^2 based classifiers beating the dummy classifier. The results were also markedly improved when the number of hand configurations was limited to 12 classes. In this limited case, the full system was evaluated with different ways of determining the best poses. In some cases, the classifier built this way beat even classifiers built around manually adjusted poses.

The classifiers were also evaluated in binary classification cases, discriminating between two hand configurations at a time. In this restricted setting, the classifiers performed adequately, commonly reaching accuracies above 80 %, and in some cases being nearly perfect. The best feature and metric combination in terms of discriminatory power turned out to be the PHOG/χ^2. Also, simple geometric features, such as the contour, worked well. This is a good piece of news because these combinations can be computed very quickly.

While the lack of training can be considered a plus, it evident that machine learning approaches could boost classifier performance significantly. Also, the current approach to probe the pose space for the best pose hypothesis is rather brutalistic and highly ineffective. Further work could reduce the amount of time needed by an order of magnitude, e.g., by employing statistical domain knowledge of actual poses present in sign languages.

Despite the seemingly modest performance of the system, the footage used for evaluating the system was very challenging, making it is very likely that with future improvements and better input, the methods evaluated in this work can serve as basis for a viable and practical system. The implementation of the methods presented in this work is freely available to the public as part of the `slmotion` package.

References

Vassilis Athitsos and Stan Sclaroff. Estimating 3D hand pose from a cluttered image. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003)*, volume 2, pages 432–439, 2003. doi: 10.1109/CVPR.2003.1211500. URL http://dx.doi.org/10.1109/CVPR.2003.1211500.

Harry Barrow, Jay Tenenbaum, Robert Bolles, and Helen Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence (IJCAI '77)*, pages 659–663, 1977.

Herbert Bay, Tinne Tuytelaars, and Luc Gool. SURF: Speeded Up Robust Features. In Aleš Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision – ECCV 2006*, volume 3951 of *Lecture Notes in Computer Science*, pages 404–417. Springer, Berlin, 2006. ISBN 978-3-540-33832-1. doi: 10.1007/11744023_32. URL http://dx.doi.org/10.1007/11744023_32.

Gunilla Borgefors. Distance transformations in digital images. *Computer Vision, Graphics, and Image Processing*, 34(3):344–371, 1986. doi: 10.1016/S0734-189X(86)80047-0. URL http://dx.doi.org/10.1016/S0734-189X(86)80047-0.

Anna Bosch, Andrew Zisserman, and Xavier Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CIVR 2007)*, pages 401–408, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-733-9. doi: 10.1145/1282280.1282340. URL http://doi.acm.org/10.1145/1282280.1282340.

Patrick Buehler, Mark Everingham, Daniel P. Huttenlocher, and Andrew Zisserman. Long term arm and hand tracking for continuous sign language TV broadcasts. In *Proceedings of the 19th British Machine Vision Conference (BMVC 2008)*, pages 1105–1114, 2008. ISBN 978-1-901725-36-0.

Patrick Buehler, Mark Everingham, Daniel P. Huttenlocher, and Andrew Zisserman. Upper body detection and tracking in extended signing sequences. *International Journal of Computer Vision*, 95(2):180–197, 2011. ISSN 0920-5691. doi: 10.1007/s11263-011-0480-9. URL http://dx.doi.org/10.1007/s11263-011-0480-9.

Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific and Statistical Computing*, 16(5):1190–1208, 1995.

John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):679–698, 1986. doi: 10.1109/TPAMI.1986.4767851. URL http://dx.doi.org/10.1109/TPAMI.1986.4767851.

Tim F. Cootes, Camillo J. Taylor, D.H. Cooper, and Jim Graham. Active shape models – their training and application. *Computer Vision and Image Understanding*, 61(1):38 – 59, 1995. ISSN 1077-3142. doi: http://dx.doi.org/10.1006/cviu.1995.1004. URL http://www.sciencedirect.com/science/article/pii/S1077314285710041.

Yuntao Cui and Juyang Weng. Appearance-based hand sign recognition from intensity image sequences. *Computer Vision and Image Understanding*, 78(2):157 – 176, 2000. ISSN 1077-3142. doi: http://dx.doi.org/10.1006/cviu.2000.0837. URL http://www.sciencedirect.com/science/article/pii/S1077314200908373.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, volume 1, pages 886–893, 2005. doi: 10.1109/CVPR.2005.177.

Martin de La Gorce, David. Fleet, and Nikolaos Paragios. Model-based 3D hand pose estimation from monocular video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(9):1793–1805, 2011. ISSN 0162-8828. doi: 10.1109/TPAMI.2011.33. URL http://dx.doi.org/10.1109/TPAMI.2011.33.

Mark Dilsizian, Polina Yanovich, Shu Wang, Carol Neidle, and Dimitris Metaxas. A new framework for sign language recognition based on 3D handshape identification and linguistic modeling. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of 9th Language Resources and Evaluation Conference (LREC 2014)*, pages 1924–1929, Reykjavík, Iceland, May 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/1138_Paper.pdf.

Finnish Association of the Deaf. Suvi, the on-line dictionary of Finnish Sign Language, 2003–2014. URL http://suvi.viittomat.net.

Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006. ISSN 0925-2312. doi: 10.1016/j.neucom.2005.12.126. URL http://dx.doi.org/10.1016/j.neucom.2005.12.126.

IEC. Recording – helical-scan digital video cassette recording system using 6,35 mm magnetic tape for consumer use (525-60, 625-50, 1125-60 and 1250-50 systems) – Part 2: SD format for 525-60 and 625-50 systems. ISO/IEC 61834-2, International Electrotechnical Commission, Geneva, Switzerland, 1998.

IEEE Computer Society. IEEE standard for floating-point arithmetic. IEEE Std 754-2008, Institute of Electrical and Electronics Engineers, New York, NY, USA, 2008. URL http://dx.doi.org/10.1109/FIEEESTD.2008.4610935.

ISO. Information technology – Programming languages – C++. ISO/IEC 14882:2011, International Organization for Standardization, Geneva, Switzerland, 2011.

Tommi Jantunen. Acceleration peaks and sonority in Finnish Sign Language syllables. In Steve Parker, editor, *The Sonority Controversy. Phonetics and Phonology 18.*, pages 347–381. Mouton De Gruyter, Berlin, 2012. ISBN 978-3-11-026152-3.

Robert E. Johnson and Scott K. Liddell. Toward a phonetic representation of hand configuration: The fingers. *Sign Language Studies*, 12(1):5–45, 2011. doi: 10.1353/sls.2011.0013. URL `http://dx.doi.org/10.1353/sls.2011.0013`.

Robert E. Johnson and Scott K. Liddell. Toward a phonetic representation of hand configuration: The thumb. *Sign Language Studies*, 12(2):316–333, 2012. doi: 10.1353/sls.2011.0020. URL `http://dx.doi.org/10.1353/sls.2011.0020`.

Atul Kanaujia, Cristian Sminchisescu, and Dimitris Metaxas. Spectral latent variable models for perceptual inference. In *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV 2007)*, pages 1–8, Oct 2007. doi: 10.1109/ICCV.2007.4408845. URL `http://dx.doi.org/10.1109/ICCV.2007.4408845`.

Matti Karppa. Model-based hand tracking methods and their applicability to estimating sign language hand configurations from video. Bachelor's thesis, Aalto University, Espoo, Finland, December 2011. In Finnish.

Matti Karppa, Tommi Jantunen, Markus Koskela, Jorma Laaksonen, and Ville Viitaniemi. Method for visualisation and analysis of hand and head movements in sign language video. In *Proceedings of the 2nd Gesture and Speech in Interaction conference (GESPIN 2011)*, Bielefeld, Germany, 2011. University of Bielefed. URL `http://coral2.spectrum.uni-bielefeld.de/gespin2011/final/Jantunen.pdf`.

Matti Karppa, Tommi Jantunen, Ville Viitaniemi, Jorma Laaksonen, Birgitta Burger, and Danny De Weerdt. Comparing computer vision analysis of signed language video with motion capture recordings. In *Proceedings of 8th Language Resources and Evaluation Conference (LREC 2012)*, pages 2421–2425, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2012/pdf/321_Paper.pdf`.

Matti Karppa, Ville Viitaniemi, Marcos Luzardo, Jorma Laaksonen, and Tommi Jantunen. SLMotion – an extensible sign language oriented video analysis tool. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of 9th Language Resources and Evaluation Conference (LREC 2014)*, Reykjavík, Iceland, May 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4. URL `http://www.lrec-conf.org/proceedings/lrec2014/pdf/209_Paper.pdf`.

Mathias Kölsch and Matthew Turk. Analysis of rotational robustness of hand detection with a Viola-Jones detector. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004)*, volume 3, pages 107–110, Aug 2004. doi: 10.1109/ICPR.2004.1334480. URL `http://dx.doi.org/10.1109/ICPR.2004.1334480`.

Ana Kuzmanic and Vlasta Zanchi. Hand shape classification using dtw and lcss as similarity measures for vision-based gesture recognition system. In *EUROCON, 2007. The International Conference on "Computer as a Tool"*, pages 264–269, 2007. doi: 10.1109/EURCON.2007.4400350.

David G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV 1999)*, volume 2, pages 1150–1157, 1999. doi: 10.1109/ICCV.1999.790410. URL `http://dx.doi.org/10.1109/ICCV.1999.790410`.

Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence (IJCAI '81)*, pages 674–679, 1981.

Naoaki Okazaki. LibLBFGS, 2002–2010. URL `http://www.chokkan.org/software/liblbfgs/`.

Eng-Jon Ong and Richard Bowden. A boosted classifier tree for hand shape detection. In *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition (FG 2004)*, pages 889–894, May 2004. doi: 10.1109/AFGR.2004.1301646. URL `http://dx.doi.org/10.1109/AFGR.2004.1301646`.

Ofir Pele and Michael Werman. A linear time histogram metric for improved sift matching. In David Forsyth, Philip Torr, and Andrew Zisserman, editors, *Computer Vision – ECCV 2008*, volume 5304 of *Lecture Notes in Computer Science*, pages 495–508. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-88689-1. doi: 10.1007/978-3-540-88690-7_37. URL `http://dx.doi.org/10.1007/978-3-540-88690-7_37`.

Ofir Pele and Michael Werman. Fast and robust Earth Mover's Distances. In *Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV 2009)*, pages 460–467, Sept 2009. doi: 10.1109/ICCV.2009.5459199. URL `http://dx.doi.org/10.1109/ICCV.2009.5459199`.

Tomas Pfister, James Charles, Mark Everingham, and Andrew Zisserman. Automatic and efficient long term arm and hand tracking for continuous sign language TV broadcasts. In *Proceedings of the British Machine Vision Conference (BMVC 2012)*, pages 4.1–4.11. BMVA Press, 2012. ISBN 1-901725-46-4. doi: 10.5244/C.26.4. URL `http://dx.doi.org/10.5244/C.26.4`.

Son Lam Phung, Abdesselam Bouzerdoum, and Douglas Chai. Skin segmentation using color pixel classification: analysis and comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(1):148–154, Jan 2005. ISSN 0162-8828. doi: 10.1109/TPAMI.2005.17. URL http://dx.doi.org/10.1109/TPAMI.2005.17.

David Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1): 37–63, 2011.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes: The Art of Scientific Computing (Third Edition)*. Cambridge University Press, New York, NY, USA, 2007. ISBN 978-0-521-88068-8. URL http://www.nrbook.com/.

Anna Puupponen. Horisontaaliset ja vertikaaliset päänliikkeet suomalaisessa viittomakielessä [in Finnish]. Master's thesis, University of Jyväskylä, Jyväskylä, Finland, 2012. URL http://urn.fi/URN:NBN:fi:jyu-201207242120.

Python Software Foundation. The Python language reference, version 2.7, 1990–2014. URL http://docs.python.org/2.7/reference/index.html.

Azriel Rosenfeld and John L. Pfaltz. Sequential operations in digital picture processing. *Journal of the ACM (JACM)*, 13(4):471–494, 1966.

Yossi Rubner, Carl Tomasi, and Leonidas J. Guibas. A metric for distributions with applications to image databases. In *Proceedings of the Sixth International Conference on Computer Vision (ICCV 1998)*, pages 59–66, Jan 1998. doi: 10.1109/ICCV.1998.710701. URL http://dx.doi.org/10.1109/ICCV.1998.710701.

Jianbo Shi and Carlo Tomasi. Good features to track. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '94)*, pages 593–600, June 1994. doi: 10.1109/ICCVW.2013.40. URL http://dx.doi.org/10.1109/ICCVW.2013.40.

Thad Starner, Joshua Weaver, and Alex Pentland. Real-time American sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375, Dec 1998. ISSN 0162-8828. doi: 10.1109/34.735811. URL http://dx.doi.org/10.1109/34.735811.

Björn Stenger, Arsanathan Thayananthan, Philip H. S. Torr, and Roberto Cipolla. Model-based hand tracking using a hierarchical bayesian filter. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1372–1384, 2006. ISSN 0162-8828. doi: 10.1109/TPAMI.2006.189. URL http://dx.doi.org/10.1109/TPAMI.2006.189.

Satoshi Suzuki and Keiichi Abe. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing*, 30(1):32–46, 1985. doi: 10.1016/0734-189X(85)90016-7. URL http://dx.doi.org/10.1016/0734-189X(85)90016-7.

Ville Viitaniemi, Matti Karppa, Jorma Laaksonen, and Tommi Jantunen. Detecting hand-head occlusions in sign language video. In *Proceedings of the 18th Scandinavian Conference on Image Analysis*, volume 7944 of *LNCS*, pages 361–372, Berlin, June 2013. Springer. doi: 10.1007/978-3-642-38886-6_35. URL http://dx.doi.org/10.1007/978-3-642-38886-6_35.

Ville Viitaniemi, Tommi Jantunen, Leena Savolainen, Matti Karppa, and Jorma Laaksonen. S-pot – a benchmark in spotting signs within continuous signing. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of 9th Language Resources and Evaluation Conference (LREC 2014)*, Reykjavík, Iceland, May 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/440_Paper.pdf.

Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, volume 1, pages 511–518. IEEE, 2001. doi: 10.1109/CVPR.2001.990517. URL http://dx.doi.org/10.1109/CVPR.2001.990517.

Marin Šarić. Libhand: A library for hand articulation, 2011. URL http://www.libhand.org/. Version 0.9.

Jingbin Wang, Vassilis Athitsos, Stan Sclaroff, and Margrit Betke. Detecting objects of variable shape structure with hidden state shape models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):477–492, 2008. ISSN 0162-8828. doi: 10.1109/TPAMI.2007.1178. URL http://dx.doi.org/10.1109/TPAMI.2007.1178.

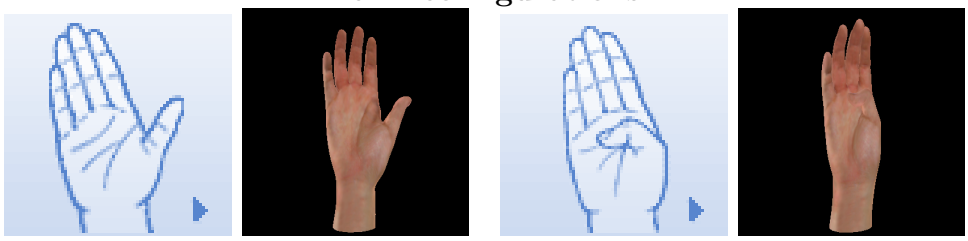
Willow Garage. OpenCV Library, 1999–2014. URL http://www.opencv.org/.

Ciyou Zhu, Richard Byrd, Jorge Nocedal, and Jose Luis Morales. Software for large-scale bound-constrained optimization, 1999–2014. URL http://users.eecs.northwestern.edu/~nocedal/lbfgsb.html.

A Suvi hand configurations and the respective renderings

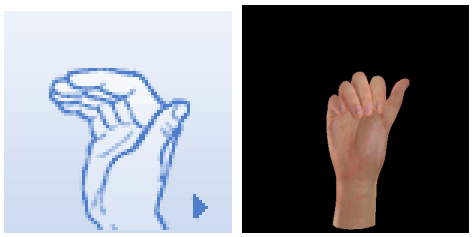
Below are the hand configurations that were used in the experiments. The Suvi model image is shown for each configuration, along with the corresponding synthetic version. The Suvi hand configuration code is also given for each configuration. Grouping of the configurations follows the one given in Suvi.

Palm configurations

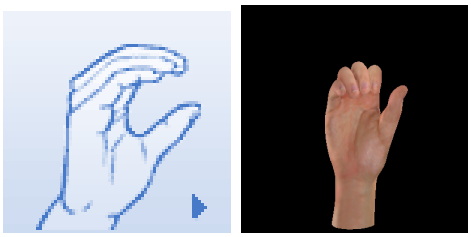


1001

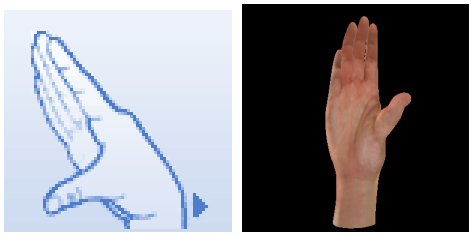
1011



1021

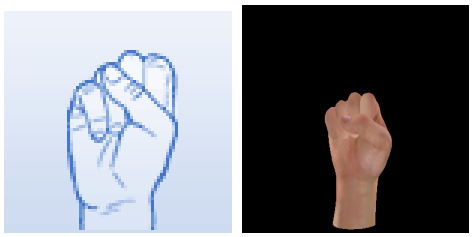


1031

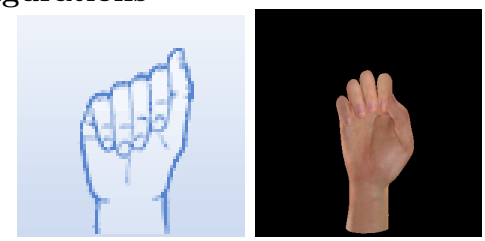


1041

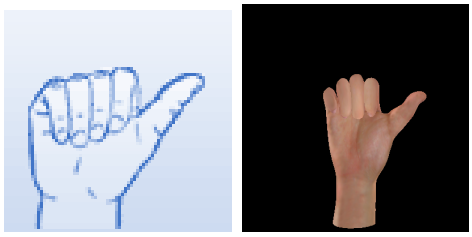
Fist configurations



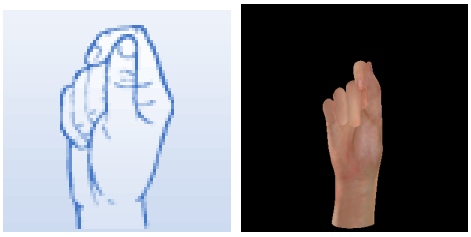
1100



1110

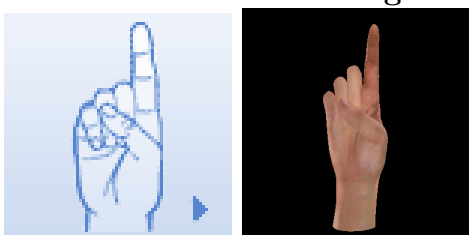


1120

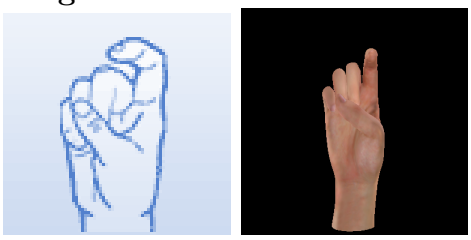


1130

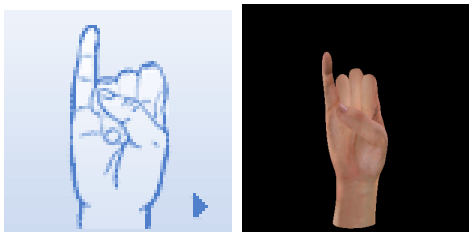
One-fingered configurations



1201

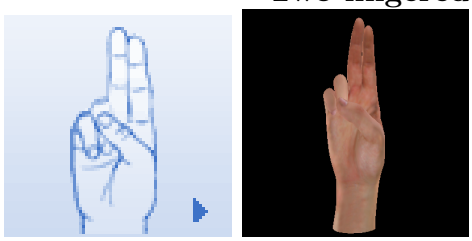


1210

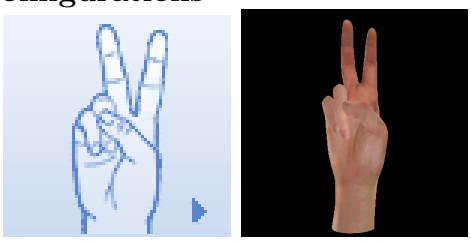


1221

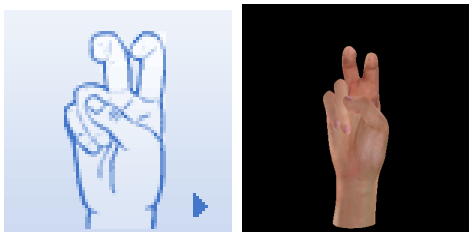
Two-fingered configurations



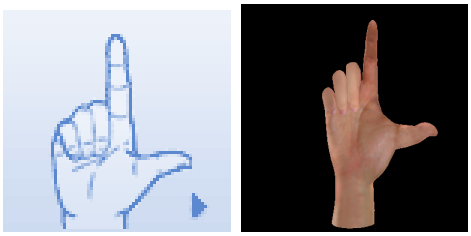
1301



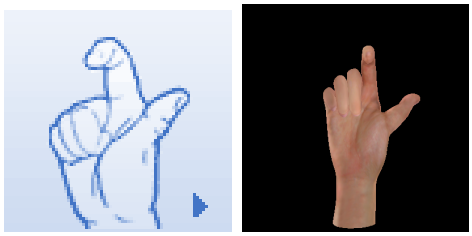
1311



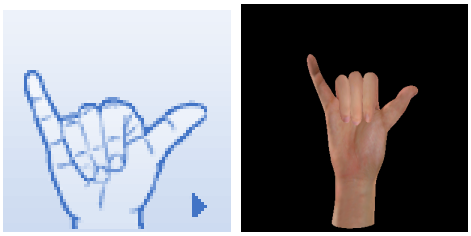
1321



1331

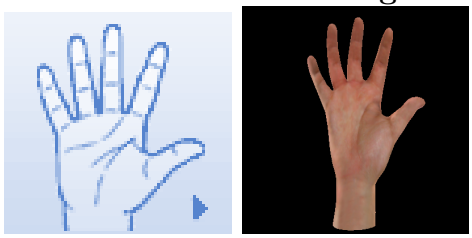


1341

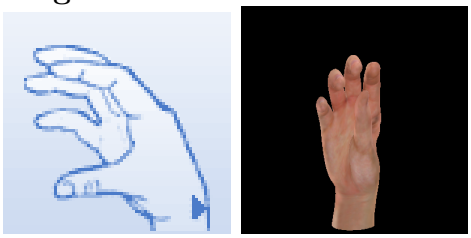


1371

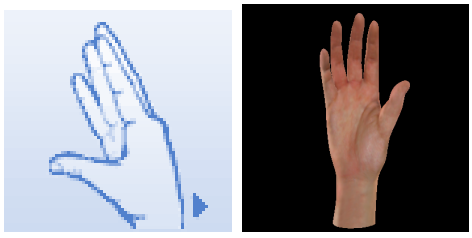
3–5-fingered configurations



1501

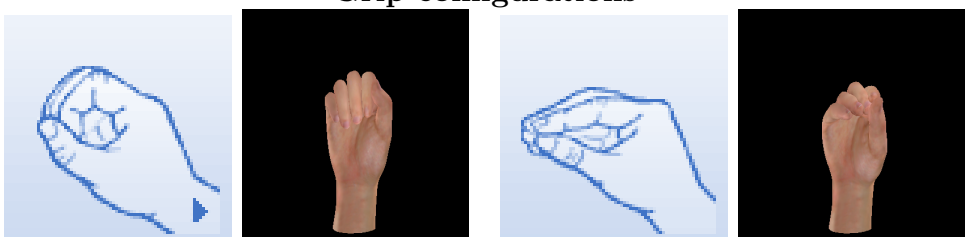


1511

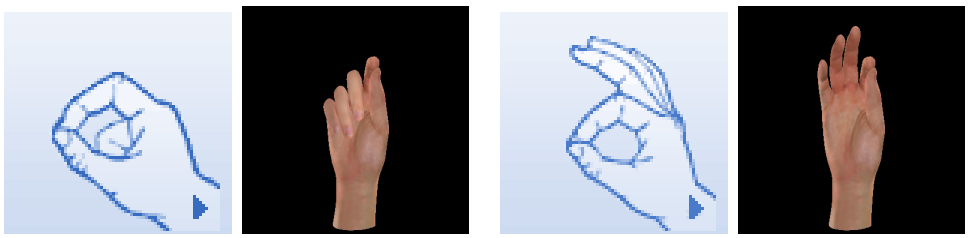


1521

Grip configurations



1601 1610



1621 1631

B Full Nearest Neighbor accuracy matrix (26 classes)

Accuracy figures for manually fitted hand poses, with respect to different discriminatory features and metrics, in the full 26-class case. Abbreviations: Gi refers to the group, the members of which have i as the second digit in their class number, TGA = Total groupwise accuracy, THOG = Trimmed HOG, Cf = Chamfer, Eu = Euclidean. The three numbers in Gi and Avg fields present the precision/recall/accuracy values, respectively. The numbers in the Avg field are averages over the Gi values.

Feature/Metric	Acc	G0	G1	G2	G3	G5	G6	Avg	TGA
Dummy	17.7 %	28.7/100 %	0/0 %	0/0 %	0/0 %	0/0 %	0/0 %	4.8/16.7 %	28.7 %
Contour/Cf	16.9 %	54.4/42.0/65.4 %	26.1/46.2/79.7 %	16.7/35.3/82.7 %	24.1/12.5/70.0 %	20.5/36.0/78.5 %	21.4/12.0/86.1 %	27.2/30.7/77.1 %	31.2 %
Canny/Cf	16.0 %	57.4/40.6/63.7 %	23.9/55.0/81.4 %	5.6/18.2/81.9 %	24.1/13.7/72.2 %	22.7/29.4/75.5 %	14.3/8.0/85.2 %	24.7/27.5/76.7 %	30.0 %
HOG/Eu	15.6 %	36.8/42.4/67.5 %	47.8/31.0/69.2 %	5.6/7.7/75.5 %	10.3/9.1/76.4 %	9.1/19.0/75.9 %	0.0/0.0/82.7 %	18.3/18.2/74.5 %	23.6 %
HOG/χ^2	16.5 %	39.7/43.5/67.9 %	45.7/32.3/70.9 %	2.8/4.8/76.8 %	17.2/13.5/76.4 %	9.1/21.1/76.8 %	7.1/3.0/81.0 %	20.3/19.7/75.0 %	24.9 %
PHOG/Eu	15.2 %	45.6/54.4/73.4 %	21.7/52.6/81.0 %	11.1/15.4/77.2 %	27.6/8.5/54.9 %	18.2/44.4/80.6 %	21.4/13.0/86.9 %	24.3/31.4/75.7 %	27.0 %
PHOG/χ^2	18.6 %	42.6/51.8/72.2 %	26.1/44.4/79.3 %	27.8/26.3/77.2 %	24.1/9.7/63.3 %	15.9/36.8/79.3 %	28.6/16.0/86.9 %	27.5/30.9/76.4 %	29.1 %
THOG/EMD	13.1 %	50.0/39.1/63.3 %	15.2/53.8/81.0 %	16.7/27.3/80.6 %	34.5/15.2/68.4 %	15.9/29.2/77.2 %	21.4/12.0/86.1 %	25.6/29.4/76.1 %	28.3 %
THOG/$\widehat{\text{EMD}}$	9.7 %	30.9/48.8/70.9 %	82.6/29.9/59.1 %	0.0/0.0/78.5 %	10.3/11.1/78.9 %	11.4/45.5/81.0 %	14.3/14.3/89.9 %	24.9/24.9/76.4 %	29.1 %
SIFT/EMD	4.6 %	14.7/32.3/66.7 %	0.0/0.0/78.5 %	5.6/11.8/79.3 %	34.5/11.0/57.8 %	29.5/21.3/66.7 %	21.4/9.4/83.1 %	17.6/14.3/72.0 %	16.0 %
SURF/EMD	4.2 %	8.8/22.2/65.0 %	6.5/10.7/71.3 %	2.8/10.0/81.4 %	51.7/12.3/48.9 %	15.9/17.1/70.0 %	21.4/33.3/92.8 %	17.9/17.6/71.6 %	14.8 %
SIFT/$\widehat{\text{EMD}}$	4.6 %	13.2/33.3/67.5 %	4.3/10.5/74.3 %	13.9/20.8/78.9 %	37.9/13.9/63.7 %	29.5/24.5/70.0 %	14.3/5.7/81.0 %	18.9/18.1/72.6 %	17.7 %
SURF/$\widehat{\text{EMD}}$	5.9 %	11.8/25.0/64.6 %	21.7/16.1/62.9 %	25.0/20.9/74.3 %	20.7/23.1/81.9 %	4.5/13.3/76.8 %	28.6/6.8/72.6 %	18.7/17.5/72.2 %	16.5 %
Majority vote	18.6 %	47.1/47.1/69.6 %	34.8/39.0/76.8 %	11.1/25.0/81.4 %	31.0/13.6/67.5 %	13.6/26.1/76.8 %	0.0/0.0/84.4 %	22.9/25.1/76.1 %	28.3 %

C Confusion matrices (26 classes)

The following are confusion matrices of different nearest-neighbor classifiers, built from manually annotated pose data. The rows correspond to the actual classes (ground truth), and columns to nearest-neighbor predictions. Diagonal elements correspond to correct classifications, and are thus in bold. Hand configuration groups have been separated with vertical and horizontal lines. Values in the same box as the bolded value constitute intra-group confusion, and values in different boxes correspond to inter-group confusion.

Contour/Chamfer:

	1001	1011	1021	1031	1041	1100	1110	1120	1130	1201	1210	1221	1301	1311	1321	1331	1341	1371	1381	1501	1511	1521	1601	1610	1621	1631
1001	**24**	3	0	1	0	0	0	0	0	0	0	1	0	1	0	0	2	0	1	2	0	2	0	0	0	5
1011	1	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0
1021	5	1	**0**	0	0	0	0	0	0	0	0	0	0	0	3	1	3	1	1	0	0	1	0	1	0	0
1031	0	1	0	**0**	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1041	1	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1	0	0	0	0	0
1100	0	3	0	0	0	**2**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
1110	0	0	0	0	0	2	**0**	1	0	0	2	0	0	0	0	0	0	0	0	0	2	0	0	2	0	0
1120	1	1	4	0	0	1	0	**0**	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
1130	0	7	0	0	0	3	0	0	**3**	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	4	0
1201	0	5	0	0	0	1	0	1	4	**4**	2	0	1	1	2	1	3	0	1	0	3	0	0	0	0	0
1210	0	1	0	0	0	2	0	0	0	0	**0**	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
1221	2	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1301	2	0	0	0	0	0	0	0	1	1	2	0	**1**	0	0	1	0	0	0	0	0	0	0	0	1	0
1311	1	4	0	0	0	0	0	0	1	0	0	0	0	**1**	1	0	0	0	0	0	0	2	0	0	0	1
1321	0	1	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0
1331	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0
1341	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	**2**	0	0	0	0	0	0	0	0	1
1371	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0
1381	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	1	0	0	0	0	0	0
1501	6	2	0	0	0	0	0	0	1	0	0	1	0	0	4	0	0	0	0	**1**	1	2	0	0	1	3
1511	2	2	0	1	1	0	0	0	0	0	0	0	0	0	3	0	3	1	2	4	**0**	0	0	0	0	1
1521	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0
1601	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	2	1	0	0	0	0	**0**	0	0	1
1610	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	1
1621	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	**0**	0
1631	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	**0**

Canny/Chamfer:

	1001	1011	1021	1031	1041	1100	1110	1120	1130	1201	1210	1221	1301	1311	1321	1331	1341	1371	1381	1501	1511	1521	1601	1610	1621	1631
1001	**22**	4	0	1	0	0	0	0	0	0	0	1	0	2	1	0	1	0	1	4	0	2	0	0	0	3
1011	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
1021	5	1	**0**	0	1	0	1	0	0	0	0	1	0	0	1	0	4	1	0	0	1	1	0	0	0	0
1031	0	0	0	**0**	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1041	2	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0
1100	1	3	0	0	0	**0**	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0
1110	0	0	0	0	1	1	**0**	1	0	0	1	0	0	0	0	0	0	0	1	0	2	0	0	2	0	0
1120	0	2	2	0	1	0	0	**2**	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
1130	0	7	0	0	1	3	0	0	**4**	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	3	0
1201	1	7	1	0	2	1	0	0	2	**1**	1	0	2	1	2	0	2	0	1	0	2	0	0	1	2	0
1210	0	2	0	0	1	1	0	0	0	0	**0**	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
1221	2	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1301	1	0	0	0	1	0	0	0	1	0	1	0	**0**	1	1	0	0	0	0	1	0	0	0	0	2	0
1311	1	5	0	0	0	0	0	0	0	0	0	0	0	**1**	1	0	0	0	0	2	0	1	0	0	0	0
1321	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	1	0	0	0	0	0	0	0	0	0
1331	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	1	0
1341	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	**1**	0	0	0	1	0	0	0	0	0
1371	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	1	0
1381	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	1	0	0	0	0	0	0
1501	3	1	0	0	0	0	0	0	0	0	1	1	0	0	8	0	0	0	0	**5**	0	1	0	0	0	2
1511	1	2	0	0	1	0	1	0	0	0	0	0	0	0	4	0	4	1	0	2	**0**	1	0	0	1	2
1521	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	1
1601	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	**0**	0	2	0
1610	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	**0**	0	0
1621	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0
1631	0	1	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**

PHOG/Euclidean:

	1001	1011	1021	1031	1041	1100	1110	1120	1130	1201	1210	1221	1301	1311	1321	1331	1341	1371	1381	1501	1511	1521	1601	1610	1621	1631
1001	**17**	2	0	1	8	0	0	0	0	0	1	0	2	2	0	3	0	0	0	5	0	1	0	0	0	0
1011	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0
1021	0	0	**0**	1	0	0	1	0	0	0	1	0	3	0	1	0	3	4	0	0	0	0	1	0	1	1
1031	0	0	0	**0**	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1041	1	0	0	0	**1**	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1100	0	0	0	1	0	**0**	1	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1
1110	0	0	0	0	0	0	**3**	0	0	0	0	0	0	2	1	0	0	1	0	0	0	1	0	0	0	1
1120	0	1	0	0	0	0	0	**2**	0	0	0	0	0	0	4	1	1	1	0	0	0	0	0	0	0	0
1130	0	1	0	2	0	0	3	0	**0**	0	4	0	2	0	6	0	1	0	0	0	0	0	0	0	1	1
1201	1	3	0	0	2	0	1	0	0	**3**	1	0	8	1	4	1	2	0	0	0	1	0	0	0	0	1
1210	0	0	0	0	0	0	0	0	1	0	**0**	0	1	0	2	0	0	0	1	0	0	0	0	0	0	0
1221	0	0	0	0	0	0	0	0	0	0	0	**0**	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1301	0	1	0	0	2	0	0	0	0	0	3	0	**1**	0	1	0	0	0	1	0	0	0	0	0	0	0
1311	0	2	0	0	1	0	0	0	0	1	2	2	0	**1**	0	0	0	0	0	0	0	0	0	0	0	2
1321	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0
1331	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0
1341	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	**0**	0	0	0	0	0	0	0	0	3
1371	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	1	0
1381	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	**0**	0	0	0	0	0	0	0
1501	2	1	0	0	2	0	0	0	1	1	0	0	2	0	2	1	1	2	1	**5**	1	0	0	0	0	0
1511	0	0	0	0	2	0	1	1	0	0	1	1	2	0	3	0	2	0	0	0	**0**	2	0	0	0	5
1521	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	1
1601	0	0	0	0	0	0	0	0	0	0	0	0	1	0	4	0	1	0	0	0	0	0	**0**	0	1	0
1610	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0
1621	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	1
1631	0	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**

PHOG/χ^2:

	1001	1011	1021	1031	1041	1100	1110	1120	1130	1201	1210	1221	1301	1311	1321	1331	1341	1371	1381	1501	1511	1521	1601	1610	1621	1631
1001	**20**	1	0	1	2	0	1	0	0	0	2	1	3	1	1	0	2	0	0	5	0	0	0	0	0	2
1011	0	**0**	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
1021	0	1	**0**	1	0	0	1	0	0	0	2	0	1	0	2	0	3	4	1	0	0	0	0	0	0	1
1031	0	0	0	**0**	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1041	1	0	0	0	**2**	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1100	0	0	0	1	0	**1**	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
1110	0	0	0	0	0	0	**2**	0	2	0	1	0	0	1	0	0	0	0	0	0	0	0	2	0	0	1
1120	0	1	1	0	0	0	0	**2**	0	0	1	0	0	1	2	0	1	1	0	0	0	0	0	0	0	0
1130	0	1	0	3	0	0	2	0	**1**	0	5	0	2	0	4	0	1	0	0	0	0	0	0	0	1	1
1201	0	3	0	0	0	0	1	1	1	**7**	2	0	5	2	2	0	2	0	0	0	2	0	0	0	0	1
1210	0	0	1	0	0	0	0	0	3	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1221	0	0	0	0	0	0	0	0	0	0	0	**0**	0	1	1	0	0	0	0	0	0	0	0	0	0	0
1301	0	0	0	0	1	0	0	0	0	1	3	0	**1**	1	1	0	0	0	1	0	0	0	0	0	0	0
1311	0	4	0	0	0	0	0	0	0	0	1	2	0	**0**	0	0	0	0	0	1	0	2	0	0	0	1
1321	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0
1331	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0
1341	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	**0**	0	0	0	0	0	0	0	0	3
1371	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	1	0
1381	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	**0**	0	0	0	0	0	0	0
1501	2	2	0	0	1	0	3	0	0	0	0	2	1	1	1	0	1	0	1	**5**	0	0	0	0	0	2
1511	0	2	0	0	1	0	1	0	0	0	1	1	0	2	3	0	2	2	0	1	**0**	1	0	0	0	3
1521	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	1
1601	0	0	0	0	0	0	0	0	0	0	0	0	1	0	4	0	1	0	0	0	0	0	**0**	0	1	0
1610	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	1
1621	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	1
1631	0	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**

Majority vote:

	1001	1011	1021	1031	1041	1100	1110	1120	1130	1201	1210	1221	1301	1311	1321	1331	1341	1371	1381	1501	1511	1521	1601	1610	1621	1631
1001	**22**	5	0	1	0	0	0	1	0	0	1	2	0	1	0	0	3	0	0	3	0	1	0	0	0	2
1011	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0
1021	1	0	**0**	1	0	0	1	0	0	0	0	1	0	0	3	0	5	1	1	0	0	1	0	1	0	1
1031	0	0	0	**0**	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
1041	1	0	0	0	**1**	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
1100	0	2	0	1	0	**2**	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
1110	0	0	0	0	0	3	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	2	0	2
1120	0	2	1	1	0	1	0	**2**	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0
1130	0	5	0	1	0	2	1	0	**5**	0	2	0	0	0	2	0	1	2	0	0	0	0	0	0	0	0
1201	0	7	0	0	0	3	0	2	3	**2**	1	1	4	0	2	0	1	0	0	0	2	0	0	1	0	0
1210	0	0	0	0	0	2	0	0	2	0	**0**	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1221	1	0	0	0	0	0	0	0	0	0	0	**0**	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1301	0	0	0	0	1	0	0	0	2	0	1	0	**1**	0	1	1	0	0	1	0	0	0	0	0	1	0
1311	0	4	0	0	0	2	0	0	0	0	0	0	0	**1**	0	0	0	0	0	1	0	1	0	0	0	2
1321	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0
1331	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0
1341	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	**1**	0	0	0	1	0	0	0	0	1
1371	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0
1381	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	1	0	0	0	0	0	0
1501	3	1	0	0	0	0	1	0	0	0	0	1	1	0	5	0	1	1	0	**5**	0	0	0	0	1	2
1511	0	1	1	0	1	1	0	1	0	0	0	1	0	0	4	0	2	2	0	1	**0**	0	0	1	0	4
1521	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	1
1601	0	0	0	0	0	0	0	0	2	0	0	0	1	0	2	0	1	1	0	0	0	0	**0**	0	0	0
1610	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	**0**	0	0
1621	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	**0**	0
1631	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	**0**

D Full Nearest Neighbor accuracy matrix (2 most common classes)

Below, the accuracies of all nearest neighbor classifiers with all fitting schemes are shown for classes 1001/1201. Rows/columns correspond to fitting schemes/discriminatory features/metrics, respectively. Abbreviations: SF/M = Scan feature/metric, Nn = none, GD = gradient descent, Dmy = dummy, Ct = contour, Cn = Canny, Cf = chamfer, Eu = Euclidean, PH = PHOG, TH = Trimmed HOG, Si = SIFT, Su = SURF, MV = majority vote

Scan feature / metric	Dmy	Ct/Cf	Cn/Cf	HOG/Eu	HOG/χ^2	PHOG/Eu	PHOG/χ^2	THOG/EMD	THOG/$\widehat{\text{EMD}}$	Si/EMD	Su/EMD	Si/$\widehat{\text{EMD}}$	Su/$\widehat{\text{EMD}}$	MV
Manual	0.59	0.89	0.80	0.73	0.75	0.73	0.82	0.83	0.62	0.52	0.58	0.52	0.37	0.72
Ct/Cf/Nn/45°	0.59	0.59	0.55	0.51	0.58	0.62	0.63	0.58	0.39	0.54	0.45	0.66	0.45	0.62
Ct/Cf/Nn/30°	0.59	0.52	0.55	0.52	0.52	0.58	0.61	0.48	0.49	0.52	0.46	0.68	0.48	0.56
Ct/Cf/Nn/22.5°	0.59	0.51	0.55	0.44	0.45	0.52	0.44	0.49	0.61	0.58	0.49	0.59	0.48	0.48
Ct/Cf/GD	0.59	0.65	0.70	0.65	0.69	0.55	0.49	0.66	0.55	0.45	0.44	0.69	0.51	0.68
Ct/Cf/L-BFGS	0.59	0.55	0.55	0.44	0.48	0.46	0.51	0.54	0.49	0.55	0.38	0.63	0.48	0.48
Cn/Cf/Nn/45°	0.59	0.58	0.52	0.56	0.56	0.46	0.54	0.58	0.54	0.44	0.42	0.63	0.49	0.54
Cn/Cf/Nn/30°	0.59	0.51	0.56	0.48	0.51	0.61	0.54	0.54	0.51	0.49	0.42	0.45	0.44	0.54
Cn/Cf/Nn/22.5°	0.59	0.55	0.51	0.55	0.55	0.52	0.58	0.55	0.59	0.54	0.44	0.58	0.49	0.58
Cn/Cf/GD	0.59	0.79	0.65	0.69	0.72	0.55	0.61	0.69	0.51	0.39	0.46	0.49	0.49	0.68
Cn/Cf/L-BFGS	0.59	0.49	0.51	0.51	0.52	0.48	0.51	0.59	0.61	0.48	0.41	0.55	0.45	0.52
HOG/χ^2/Nn/45°	0.59	0.56	0.56	0.52	0.56	0.61	0.59	0.58	0.46	0.52	0.48	0.45	0.32	0.56
HOG/χ^2/Nn/30°	0.59	0.54	0.55	0.55	0.61	0.59	0.59	0.56	0.44	0.51	0.54	0.58	0.39	0.55
HOG/χ^2/Nn/22.5°	0.59	0.46	0.56	0.49	0.55	0.45	0.48	0.44	0.41	0.52	0.59	0.52	0.52	0.48
HOG/χ^2/GD	0.59	0.59	0.56	0.69	0.69	0.56	0.63	0.65	0.55	0.49	0.51	0.58	0.41	0.58
HOG/χ^2/L-BFGS	0.59	0.52	0.51	0.52	0.56	0.52	0.56	0.52	0.48	0.55	0.52	0.66	0.45	0.52
HOG/Eu/Nn/45°	0.59	0.52	0.56	0.54	0.55	0.56	0.55	0.56	0.44	0.48	0.51	0.45	0.34	0.52
HOG/Eu/Nn/30°	0.59	0.52	0.58	0.55	0.59	0.55	0.59	0.55	0.42	0.51	0.55	0.55	0.38	0.55
HOG/Eu/Nn/22.5°	0.59	0.44	0.52	0.48	0.54	0.48	0.51	0.44	0.38	0.48	0.54	0.54	0.55	0.46
HOG/Eu/GD	0.59	0.68	0.72	0.77	0.77	0.63	0.65	0.65	0.62	0.54	0.52	0.56	0.37	0.72
HOG/Eu/L-BFGS	0.59	0.44	0.42	0.41	0.49	0.38	0.39	0.41	0.41	0.48	0.45	0.58	0.38	0.42
PHOG/χ^2/Nn/45°	0.59	0.58	0.61	0.58	0.59	0.63	0.63	0.55	0.52	0.46	0.42	0.62	0.46	0.59
PHOG/χ^2/Nn/30°	0.59	0.49	0.54	0.49	0.52	0.65	0.58	0.42	0.49	0.45	0.55	0.61	0.51	0.59
PHOG/χ^2/Nn/22.5°	0.59	0.58	0.55	0.59	0.56	0.52	0.56	0.52	0.59	0.52	0.48	0.59	0.48	0.61
PHOG/χ^2/GD	0.59	0.49	0.49	0.52	0.49	0.55	0.56	0.54	0.48	0.51	0.49	0.65	0.46	0.54
PHOG/χ^2/L-BFGS	0.59	0.51	0.56	0.55	0.54	0.63	0.61	0.46	0.54	0.45	0.46	0.54	0.51	0.51
PHOG/Eu/Nn/45°	0.60	0.46	0.49	0.46	0.46	0.51	0.53	0.42	0.56	0.47	0.44	0.51	0.42	0.51
PHOG/Eu/Nn/30°	0.60	0.28	0.39	0.37	0.37	0.60	0.42	0.37	0.44	0.44	0.60	0.58	0.49	0.42
PHOG/Eu/Nn/22.5°	0.56	0.51	0.48	0.51	0.51	0.57	0.56	0.51	0.49	0.54	0.61	0.64	0.43	0.57
PHOG/Eu/GD	0.57	0.40	0.52	0.46	0.46	0.51	0.49	0.44	0.44	0.51	0.41	0.54	0.43	0.51
PHOG/Eu/L-BFGS	0.59	0.49	0.54	0.46	0.49	0.55	0.59	0.52	0.58	0.48	0.46	0.51	0.39	0.52
PHOG4/χ^2/Nn/45°	0.59	0.59	0.58	0.51	0.49	0.63	0.58	0.54	0.49	0.44	0.49	0.51	0.41	0.51
PHOG4/χ^2/Nn/30°	0.59	0.48	0.54	0.46	0.48	0.65	0.52	0.52	0.51	0.54	0.48	0.56	0.52	0.55
PHOG4/χ^2/Nn/22.5°	0.59	0.61	0.55	0.52	0.48	0.58	0.56	0.54	0.52	0.62	0.62	0.56	0.62	0.61
PHOG4/χ^2/GD	0.59	0.55	0.56	0.48	0.49	0.59	0.56	0.49	0.51	0.55	0.55	0.55	0.45	0.49
PHOG4/χ^2/L-BFGS	0.59	0.61	0.55	0.48	0.44	0.62	0.55	0.49	0.44	0.38	0.51	0.49	0.46	0.49
PHOG4/Eu/Nn/45°	0.59	0.60	0.60	0.48	0.52	0.60	0.59	0.50	0.50	0.48	0.50	0.53	0.48	0.53
PHOG4/Eu/Nn/30°	0.59	0.45	0.50	0.50	0.50	0.62	0.55	0.54	0.52	0.52	0.52	0.50	0.46	0.55
PHOG4/Eu/Nn/22.5°	0.59	0.48	0.52	0.46	0.46	0.57	0.55	0.54	0.52	0.55	0.57	0.62	0.50	0.52
PHOG4/Eu/GD	0.59	0.57	0.54	0.48	0.46	0.54	0.54	0.55	0.54	0.55	0.54	0.59	0.50	0.54
PHOG4/Eu/L-BFGS	0.59	0.58	0.63	0.48	0.51	0.63	0.63	0.54	0.52	0.46	0.51	0.54	0.44	0.55
PHOG5/χ^2/Nn/45°	0.59	0.59	0.56	0.54	0.54	0.59	0.55	0.56	0.52	0.39	0.49	0.48	0.51	0.51
PHOG5/χ^2/Nn/30°	0.59	0.51	0.55	0.55	0.54	0.56	0.58	0.58	0.54	0.54	0.54	0.65	0.56	0.59
PHOG5/χ^2/Nn/22.5°	0.59	0.54	0.52	0.54	0.52	0.55	0.51	0.52	0.49	0.54	0.61	0.58	0.62	0.58
PHOG5/χ^2/GD	0.59	0.49	0.51	0.52	0.55	0.55	0.55	0.58	0.56	0.55	0.58	0.59	0.42	0.52
PHOG5/χ^2/L-BFGS	0.59	0.59	0.55	0.55	0.55	0.55	0.56	0.58	0.52	0.42	0.54	0.46	0.38	0.55
TH/10/EMD/Nn/45°	0.59	0.52	0.55	0.49	0.46	0.49	0.52	0.54	0.52	0.51	0.41	0.51	0.42	0.49
TH/10/EMD/Nn/30°	0.59	0.62	0.61	0.59	0.59	0.49	0.59	0.56	0.52	0.48	0.48	0.58	0.52	0.63
TH/10/EMD/Nn/22.5°	0.59	0.54	0.54	0.48	0.48	0.52	0.54	0.52	0.48	0.45	0.52	0.62	0.46	0.55
TH/10/EMD/GD	0.59	0.56	0.51	0.56	0.55	0.55	0.55	0.62	0.56	0.39	0.45	0.58	0.51	0.56
TH/10/EMD/L-BFGS	0.59	0.59	0.63	0.58	0.56	0.61	0.65	0.62	0.44	0.52	0.51	0.48	0.37	0.59
TH/10/$\widehat{\text{EMD}}$/Nn/45°	0.59	0.52	0.55	0.56	0.55	0.54	0.59	0.55	0.54	0.45	0.39	0.52	0.45	0.55
TH/10/$\widehat{\text{EMD}}$/Nn/30°	0.59	0.55	0.56	0.48	0.48	0.41	0.49	0.54	0.54	0.44	0.44	0.45	0.51	0.48
TH/10/$\widehat{\text{EMD}}$/Nn/22.5°	0.59	0.56	0.52	0.45	0.49	0.51	0.54	0.59	0.46	0.45	0.49	0.62	0.51	0.51
TH/10/$\widehat{\text{EMD}}$/GD	0.59	0.44	0.45	0.42	0.48	0.41	0.38	0.54	0.49	0.42	0.34	0.51	0.48	0.38
TH/10/$\widehat{\text{EMD}}$/L-BFGS	0.59	0.54	0.56	0.54	0.49	0.56	0.56	0.51	0.48	0.39	0.49	0.58	0.44	0.54
TH/50/EMD/Nn/45°	0.59	0.52	0.49	0.46	0.46	0.49	0.49	0.58	0.49	0.59	0.45	0.68	0.42	0.49
TH/50/EMD/Nn/30°	0.59	0.49	0.48	0.58	0.55	0.51	0.44	0.54	0.48	0.42	0.52	0.54	0.44	0.46
TH/50/EMD/Nn/22.5°	0.59	0.55	0.54	0.58	0.59	0.61	0.63	0.65	0.61	0.54	0.45	0.56	0.46	0.62
TH/50/EMD/GD	0.59	0.69	0.69	0.68	0.69	0.52	0.56	0.63	0.44	0.32	0.44	0.52	0.48	0.59
TH/50/EMD/L-BFGS	0.59	0.48	0.46	0.51	0.51	0.49	0.48	0.51	0.54	0.52	0.45	0.59	0.61	0.51
TH/50/$\widehat{\text{EMD}}$/Nn/45°	0.59	0.51	0.49	0.48	0.46	0.44	0.48	0.48	0.46	0.51	0.44	0.61	0.39	0.44
TH/50/$\widehat{\text{EMD}}$/Nn/30°	0.59	0.46	0.46	0.58	0.55	0.51	0.51	0.48	0.54	0.46	0.46	0.46	0.48	0.51
TH/50/$\widehat{\text{EMD}}$/Nn/22.5°	0.59	0.56	0.55	0.49	0.52	0.63	0.56	0.65	0.51	0.59	0.48	0.48	0.49	0.55
TH/50/$\widehat{\text{EMD}}$/GD	0.59	0.69	0.69	0.66	0.61	0.68	0.69	0.59	0.48	0.49	0.54	0.55	0.45	0.62
TH/50/$\widehat{\text{EMD}}$/L-BFGS	0.59	0.42	0.39	0.41	0.41	0.54	0.54	0.49	0.46	0.54	0.52	0.48	0.51	0.45
TH/100/EMD/Nn/45°	0.59	0.52	0.52	0.42	0.44	0.51	0.55	0.63	0.42	0.49	0.54	0.59	0.39	0.49
TH/100/EMD/Nn/30°	0.59	0.49	0.48	0.48	0.48	0.62	0.58	0.59	0.42	0.58	0.51	0.62	0.46	0.49
TH/100/EMD/Nn/22.5°	0.59	0.45	0.49	0.41	0.41	0.62	0.62	0.59	0.38	0.55	0.58	0.66	0.54	0.54
TH/100/EMD/GD	0.59	0.65	0.66	0.70	0.70	0.58	0.65	0.63	0.51	0.44	0.56	0.63	0.46	0.62
TH/100/EMD/L-BFGS	0.59	0.51	0.49	0.42	0.42	0.48	0.44	0.51	0.37	0.51	0.52	0.59	0.51	0.44

E Full Nearest Neighbor accuracy matrices (12 classes)

Below are the full nearest-neighbor accuracy matrices for a 12-class classification task, including groupwise accuracies, analogous to those in Appendix B. Each matrix represents a different way of determining the six degrees of freedom, or the pose. As before, each row corresponds to a nearest-neighbor classifier formed on the basis of a different discriminatory feature/metric combination.

Abbreviations: Gi refers to the group, the members of which have i as the second digit in their class number, TGA = Total groupwise accuracy, THOG = Trimmed HOG, Cf = Chamfer, Eu = Euclidean. The three numbers in Gi and Avg fields present the precision/recall/accuracy values, respectively. The numbers in the Avg field are averages over the Gi values.

E.1 Manually adjusted poses

Feature/Metric	Acc	G0	G1	G2	G3	G5	G6	Avg	TGA
Dummy	26.2 %	100.0/29.4/29.4 %	0.0/0.0/81.2 %	0.0/0.0/81.9 %	0.0/0.0/84.4 %	0.0/0.0/86.2 %	0.0/0.0/95.6 %	16.7/4.9/76.5 %	29.4 %
Contour/Cf	30.0 %	70.2/47.8/68.8 %	50.0/60.0/84.4 %	17.2/55.6/82.5 %	32.0/22.9/72.5 %	18.2/19.0/78.1 %	0.0/0.0/95.0 %	31.3/34.2/80.2 %	40.6 %
Canny/Cf	28.1 %	68.1/45.1/66.2 %	40.0/50.0/81.2 %	6.9/28.6/80.0 %	28.0/21.9/73.1 %	18.2/18.2/77.5 %	28.6/50.0/95.6 %	31.6/35.6/79.0 %	36.9 %
HOG/Eu	27.5 %	70.2/57.9/76.2 %	63.3/34.5/70.6 %	6.9/66.7/82.5 %	16.0/22.2/78.1 %	4.5/8.3/80.0 %	14.3/6.7/87.5 %	29.2/32.7/79.2 %	37.5 %
HOG/χ^2	26.9 %	68.1/55.2/74.4 %	66.7/36.4/71.9 %	6.9/66.7/82.5 %	24.0/25.0/76.9 %	0.0/0.0/80.0 %	0.0/0.0/89.4 %	27.6/30.5/79.2 %	37.5 %
PHOG/Eu	23.8 %	61.7/52.7/72.5 %	26.7/42.1/79.4 %	10.3/27.3/78.8 %	36.0/15.0/58.1 %	18.2/44.4/85.6 %	0.0/0.0/91.9 %	25.5/30.3/77.7 %	33.1 %
PHOG/χ^2	26.9 %	66.0/58.5/76.2 %	30.0/42.9/79.4 %	24.1/41.2/80.0 %	32.0/16.7/64.4 %	13.6/37.5/85.0 %	14.3/7.7/88.8 %	30.0/34.1/79.0 %	36.9 %
THOG/EMD	25.6 %	68.1/40.5/61.3 %	13.3/44.4/80.6 %	27.6/42.1/80.0 %	28.0/17.1/67.5 %	9.1/22.2/83.1 %	0.0/0.0/93.8 %	24.3/27.7/77.7 %	33.1 %
THOG/$\widehat{\text{EMD}}$	20.6 %	44.7/60.0/75.0 %	80.0/28.2/58.1 %	0.0/0.0/80.0 %	8.0/13.3/77.5 %	22.7/71.4/88.1 %	0.0/0.0/86.2 %	25.9/28.8/77.5 %	32.5 %
SIFT/EMD	6.2 %	21.3/28.6/61.3 %	0.0/0.0/75.6 %	3.4/10.0/76.9 %	36.0/17.6/63.7 %	27.3/12.2/63.1 %	0.0/0.0/91.9 %	14.7/11.4/72.1 %	16.2 %
SURF/EMD	8.8 %	25.5/40.0/66.9 %	0.0/0.0/75.0 %	3.4/16.7/79.4 %	60.0/14.6/38.8 %	4.5/9.1/80.6 %	0.0/0.0/95.6 %	15.6/13.4/72.7 %	18.1 %
SIFT/$\widehat{\text{EMD}}$	11.2 %	23.4/34.4/64.4 %	20.0/26.1/74.4 %	13.8/25.0/76.9 %	44.0/23.4/68.8 %	22.7/14.7/71.2 %	0.0/0.0/90.6 %	20.7/20.6/74.4 %	23.1 %
SURF/$\widehat{\text{EMD}}$	8.1 %	27.7/29.5/59.4 %	23.3/13.5/57.5 %	3.4/7.1/74.4 %	24.0/31.6/80.0 %	0.0/0.0/80.6 %	14.3/4.5/83.1 %	15.5/14.4/72.5 %	17.5 %
Majority vote	27.5 %	66.0/53.4/73.1 %	56.7/41.5/76.9 %	6.9/66.7/82.5 %	32.0/20.5/70.0 %	4.5/7.7/79.4 %	0.0/0.0/91.9 %	27.7/31.6/79.0 %	36.9 %

E.2 Contour/Chamfer/None (45°)

Feature/Metric	Acc	G0	G1	G2	G3	G5	G6	Avg	TGA
Dummy	26.2 %	100.0/29.4/29.4 %	0.0/0.0/81.2 %	0.0/0.0/81.9 %	0.0/0.0/84.4 %	0.0/0.0/86.2 %	0.0/0.0/95.6 %	16.7/4.9/76.5 %	29.4 %
Contour/Cf	13.8 %	80.9/28.1/33.8 %	13.3/28.6/77.5 %	3.4/25.0/80.6 %	4.0/16.7/81.9 %	0.0/0.0/86.2 %	0.0/0.0/95.0 %	16.9/16.4/75.8 %	27.5 %
Canny/Cf	10.0 %	83.0/28.3/33.1 %	3.3/20.0/79.4 %	0.0/0.0/80.6 %	8.0/16.7/79.4 %	0.0/0.0/84.4 %	0.0/0.0/95.6 %	15.7/10.8/75.4 %	26.2 %
HOG/Eu	9.4 %	63.8/29.4/44.4 %	23.3/23.3/71.2 %	3.4/25.0/80.6 %	8.0/13.3/77.5 %	0.0/0.0/83.8 %	0.0/0.0/92.5 %	16.4/15.2/75.0 %	25.0 %
HOG/χ^2	10.0 %	72.3/31.8/46.2 %	20.0/19.4/69.4 %	3.4/33.3/81.2 %	8.0/16.7/79.4 %	0.0/0.0/85.6 %	0.0/0.0/91.9 %	17.3/16.9/75.6 %	26.9 %
PHOG/Eu	14.4 %	55.3/36.6/58.8 %	10.0/50.0/81.2 %	20.7/42.9/80.6 %	36.0/21.4/69.4 %	4.5/5.6/76.2 %	0.0/0.0/90.0 %	21.1/26.1/76.0 %	28.1 %
PHOG/χ^2	14.4 %	66.0/36.5/56.2 %	6.7/40.0/80.6 %	13.8/28.6/78.1 %	32.0/17.4/65.6 %	0.0/0.0/81.9 %	0.0/0.0/93.8 %	19.7/20.4/76.0 %	28.1 %
THOG/EMD	11.9 %	61.7/28.4/43.1 %	13.3/18.2/72.5 %	3.4/16.7/79.4 %	4.0/5.6/74.4 %	4.5/20.0/84.4 %	0.0/0.0/91.2 %	14.5/14.8/74.2 %	22.5 %
THOG/$\widehat{\text{EMD}}$	8.1 %	61.7/28.7/43.8 %	6.7/9.1/70.0 %	3.4/16.7/79.4 %	24.0/27.3/78.1 %	0.0/0.0/84.4 %	14.3/16.7/93.1 %	18.4/16.4/74.8 %	24.4 %
SIFT/EMD	10.6 %	19.1/32.1/64.4 %	3.3/100.0/81.9 %	3.4/16.7/79.4 %	56.0/17.5/51.9 %	22.7/11.6/65.6 %	0.0/0.0/94.4 %	17.4/29.7/72.9 %	18.8 %
SURF/EMD	5.0 %	8.5/44.4/70.0 %	13.3/50.0/81.2 %	0.0/0.0/76.9 %	72.0/17.1/41.2 %	9.1/8.0/73.1 %	0.0/0.0/92.5 %	17.2/19.9/72.5 %	17.5 %
SIFT/$\widehat{\text{EMD}}$	12.5 %	34.0/47.1/69.4 %	13.3/25.0/76.2 %	17.2/33.3/78.8 %	36.0/16.4/61.3 %	18.2/10.8/68.1 %	0.0/0.0/93.8 %	19.8/22.1/74.6 %	23.8 %
SURF/$\widehat{\text{EMD}}$	8.8 %	21.3/37.0/66.2 %	13.3/19.0/73.1 %	13.8/14.8/70.0 %	52.0/33.3/76.2 %	0.0/0.0/75.6 %	28.6/6.9/80.0 %	21.5/18.5/73.5 %	20.6 %
Majority vote	12.5 %	72.3/33.0/48.8 %	20.0/30.0/76.2 %	6.9/50.0/81.9 %	24.0/23.1/75.6 %	0.0/0.0/84.4 %	0.0/0.0/93.1 %	20.5/22.7/76.7 %	30.0 %

E.3 Contour/Chamfer/None (30°)

Feature/Metric	Acc	G0	G1	G2	G3	G5	G6	Avg	TGA
Dummy	26.2 %	100.0/29.4/29.4 %	0.0/0.0/81.2 %	0.0/0.0/81.9 %	0.0/0.0/84.4 %	0.0/0.0/86.2 %	0.0/0.0/95.6 %	16.7/4.9/76.5 %	29.4 %
Contour/Cf	13.1 %	74.5/27.3/34.4 %	16.7/31.2/77.5 %	0.0/0.0/81.2 %	8.0/16.7/79.4 %	0.0/0.0/85.6 %	0.0/0.0/94.4 %	16.5/12.5/75.4 %	26.2 %
Canny/Cf	8.1 %	78.7/27.8/33.8 %	6.7/33.3/80.0 %	0.0/0.0/81.9 %	12.0/17.6/77.5 %	0.0/0.0/85.0 %	0.0/0.0/94.4 %	16.2/13.1/75.4 %	26.2 %
HOG/Eu	11.9 %	63.8/27.5/40.0 %	23.3/31.8/76.2 %	6.9/28.6/80.0 %	12.0/17.6/77.5 %	0.0/0.0/85.6 %	0.0/0.0/93.1 %	17.7/17.6/75.4 %	26.2 %
HOG/χ^2	15.0 %	68.1/29.1/41.9 %	33.3/37.0/76.9 %	3.4/16.7/79.4 %	12.0/25.0/80.6 %	0.0/0.0/85.6 %	0.0/0.0/93.1 %	19.5/18.0/76.2 %	28.7 %
PHOG/Eu	11.9 %	46.8/36.7/60.6 %	3.3/33.3/80.6 %	17.2/33.3/78.8 %	52.0/20.6/61.3 %	0.0/0.0/75.0 %	0.0/0.0/95.0 %	19.9/20.7/75.2 %	25.6 %
PHOG/χ^2	14.4 %	70.2/40.2/60.6 %	3.3/25.0/80.0 %	6.9/20.0/78.1 %	48.0/22.2/65.6 %	0.0/0.0/81.2 %	0.0/0.0/94.4 %	21.4/17.9/76.7 %	30.0 %
THOG/EMD	13.1 %	46.8/28.2/49.4 %	26.7/28.6/73.8 %	0.0/0.0/76.2 %	36.0/25.0/73.1 %	0.0/0.0/84.4 %	0.0/0.0/91.9 %	18.2/13.6/74.8 %	24.4 %
THOG/$\widehat{\text{EMD}}$	8.8 %	61.7/32.2/50.6 %	16.7/13.2/63.7 %	0.0/0.0/79.4 %	8.0/9.5/73.8 %	0.0/0.0/85.6 %	0.0/0.0/91.9 %	14.4/9.2/74.2 %	22.5 %
SIFT/EMD	11.2 %	31.9/44.1/68.1 %	0.0/0.0/79.4 %	3.4/7.7/75.0 %	68.0/21.5/56.2 %	22.7/16.7/73.8 %	0.0/0.0/95.0 %	21.0/15.0/74.6 %	23.8 %
SURF/EMD	17.5 %	29.8/45.2/68.8 %	0.0/0.0/78.8 %	27.6/26.7/73.1 %	56.0/23.0/63.7 %	22.7/17.2/74.4 %	0.0/0.0/92.5 %	22.7/18.7/75.2 %	25.6 %
SIFT/$\widehat{\text{EMD}}$	12.5 %	31.9/42.9/67.5 %	3.3/6.2/72.5 %	6.9/12.5/74.4 %	40.0/17.5/61.3 %	27.3/19.4/74.4 %	14.3/20.0/93.8 %	20.6/19.8/74.0 %	21.9 %
SURF/$\widehat{\text{EMD}}$	8.8 %	25.5/31.6/61.9 %	10.0/15.8/73.1 %	20.7/14.6/63.7 %	24.0/17.1/70.0 %	4.5/11.1/81.9 %	0.0/0.0/84.4 %	14.1/15.0/72.5 %	17.5 %
Majority vote	14.4 %	72.3/31.8/46.2 %	23.3/41.2/79.4 %	6.9/33.3/80.6 %	20.0/16.7/71.9 %	0.0/0.0/86.2 %	0.0/0.0/95.6 %	20.4/20.5/76.7 %	30.0 %

E.4 Contour/Chamfer/None (22.5°)

Feature/Metric	Acc	G0	G1	G2	G3	G5	G6	Avg	TGA
Dummy	26.2 %	100.0/29.4/29.4 %	0.0/0.0/81.2 %	0.0/0.0/81.9 %	0.0/0.0/84.4 %	0.0/0.0/86.2 %	0.0/0.0/95.6 %	16.7/4.9/76.5 %	29.4 %
Contour/Cf	14.4 %	76.6/29.3/38.8 %	10.0/37.5/80.0 %	0.0/0.0/80.0 %	24.0/24.0/76.2 %	0.0/0.0/85.6 %	0.0/0.0/95.6 %	18.4/15.1/76.0 %	28.1 %
Canny/Cf	8.8 %	91.5/30.3/35.6 %	3.3/100.0/81.9 %	0.0/0.0/80.0 %	8.0/14.3/78.1 %	0.0/0.0/86.2 %	0.0/0.0/95.6 %	17.1/24.1/76.2 %	28.7 %
HOG/Eu	13.1 %	68.1/29.4/42.5 %	16.7/23.8/74.4 %	0.0/0.0/79.4 %	8.0/10.5/75.0 %	4.5/33.3/85.6 %	0.0/0.0/93.1 %	16.2/16.2/75.0 %	25.0 %
HOG/χ^2	13.8 %	68.1/29.4/42.5 %	20.0/27.3/75.0 %	0.0/0.0/79.4 %	4.0/5.0/73.1 %	4.5/33.3/85.6 %	0.0/0.0/94.4 %	16.1/15.8/75.0 %	25.0 %
PHOG/Eu	16.2 %	57.4/42.9/65.0 %	6.7/28.6/79.4 %	6.9/16.7/76.9 %	56.0/23.3/64.4 %	4.5/6.2/77.5 %	0.0/0.0/94.4 %	21.9/19.6/76.2 %	28.7 %
PHOG/χ^2	15.0 %	61.7/37.2/58.1 %	3.3/12.5/77.5 %	3.4/12.5/78.1 %	56.0/23.3/64.4 %	4.5/20.0/84.4 %	0.0/0.0/95.0 %	21.5/17.6/76.2 %	28.7 %
THOG/EMD	18.8 %	74.5/38.0/56.9 %	16.7/27.8/76.2 %	6.9/40.0/81.2 %	40.0/23.8/70.6 %	4.5/100.0/86.9 %	0.0/0.0/94.4 %	23.8/38.3/77.7 %	33.1 %
THOG/$\widehat{\text{EMD}}$	8.8 %	57.4/31.8/51.2 %	26.7/22.2/68.8 %	0.0/0.0/78.8 %	28.0/29.2/78.1 %	0.0/0.0/84.4 %	0.0/0.0/91.2 %	18.7/13.9/75.4 %	26.2 %
SIFT/EMD	6.2 %	12.8/17.1/56.2 %	0.0/0.0/77.5 %	10.3/21.4/76.9 %	44.0/19.6/63.1 %	9.1/4.3/60.0 %	14.3/33.3/95.0 %	15.1/16.0/71.5 %	14.4 %
SURF/EMD	7.5 %	14.9/43.8/69.4 %	0.0/0.0/79.4 %	10.3/14.3/72.5 %	68.0/17.3/44.4 %	13.6/18.8/80.0 %	0.0/0.0/91.9 %	17.8/15.7/72.9 %	18.8 %
SIFT/$\widehat{\text{EMD}}$	11.2 %	23.4/28.2/60.0 %	16.7/31.2/77.5 %	3.4/12.5/78.1 %	40.0/21.7/68.1 %	27.3/15.0/68.8 %	0.0/0.0/88.8 %	18.5/18.1/73.5 %	20.6 %
SURF/$\widehat{\text{EMD}}$	5.6 %	34.0/37.2/63.7 %	16.7/11.9/61.3 %	3.4/7.7/75.0 %	36.0/20.0/67.5 %	9.1/40.0/85.6 %	0.0/0.0/88.1 %	16.5/19.5/73.5 %	20.6 %
Majority vote	14.4 %	74.5/33.0/48.1 %	16.7/45.5/80.6 %	0.0/0.0/80.0 %	28.0/18.9/70.0 %	4.5/50.0/86.2 %	0.0/0.0/95.0 %	20.6/24.6/76.7 %	30.0 %

E.5 Contour/Chamfer/Gradient descent

Feature/Metric	Acc	G0	G1	G2	G3	G5	G6	Avg	TGA
Dummy	26.2 %	100.0/29.4/29.4 %	0.0/0.0/81.2 %	0.0/0.0/81.9 %	0.0/0.0/84.4 %	0.0/0.0/86.2 %	0.0/0.0/95.6 %	16.7/4.9/76.5 %	29.4 %
Contour/Cf	21.2 %	51.1/38.7/61.9 %	13.3/26.7/76.9 %	37.9/61.1/84.4 %	24.0/13.0/63.1 %	22.7/45.5/85.6 %	28.6/25.0/93.1 %	29.6/35.0/77.5 %	32.5 %
Canny/Cf	19.4 %	48.9/39.7/63.1 %	13.3/33.3/78.8 %	27.6/50.0/81.9 %	48.0/27.9/72.5 %	36.4/36.4/82.5 %	0.0/0.0/90.0 %	29.0/31.2/78.1 %	34.4 %
HOG/Eu	20.0 %	46.8/53.7/72.5 %	20.0/26.1/74.4 %	27.6/36.4/78.1 %	44.0/25.6/71.2 %	22.7/41.7/85.0 %	14.3/5.3/85.0 %	29.2/31.4/77.7 %	33.1 %
HOG/χ^2	19.4 %	51.1/49.0/70.0 %	20.0/28.6/75.6 %	27.6/42.1/80.0 %	36.0/23.1/71.2 %	22.7/45.5/85.6 %	14.3/4.8/83.8 %	28.6/32.2/77.7 %	33.1 %
PHOG/Eu	9.4 %	29.8/36.8/64.4 %	26.7/29.6/74.4 %	10.3/16.7/74.4 %	44.0/21.6/66.2 %	4.5/8.3/80.0 %	0.0/0.0/86.9 %	19.2/18.8/74.4 %	23.1 %
PHOG/χ^2	13.8 %	34.0/34.0/61.3 %	26.7/32.0/75.6 %	6.9/14.3/75.6 %	52.0/24.5/67.5 %	13.6/20.0/80.6 %	0.0/0.0/91.9 %	22.2/20.8/75.4 %	26.2 %
THOG/EMD	15.6 %	38.3/45.0/68.1 %	16.7/19.2/71.2 %	13.8/28.6/78.1 %	60.0/27.3/68.8 %	18.2/28.6/82.5 %	0.0/0.0/88.8 %	24.5/24.8/76.2 %	28.7 %
THOG/$\widehat{\text{EMD}}$	10.6 %	23.4/21.2/51.9 %	10.0/15.8/73.1 %	6.9/28.6/80.0 %	40.0/23.3/70.0 %	18.2/12.9/71.9 %	14.3/12.5/91.9 %	18.8/19.0/73.1 %	19.4 %
SIFT/EMD	7.5 %	21.3/25.6/58.8 %	3.3/10.0/76.2 %	6.9/18.2/77.5 %	36.0/14.5/56.9 %	27.3/18.2/73.1 %	0.0/0.0/92.5 %	15.8/14.4/72.5 %	17.5 %
SURF/EMD	11.2 %	19.1/28.1/61.9 %	13.3/44.4/80.6 %	17.2/29.4/77.5 %	52.0/18.6/56.9 %	9.1/8.0/73.1 %	0.0/0.0/91.2 %	18.5/21.4/73.5 %	20.6 %
SIFT/$\widehat{\text{EMD}}$	8.8 %	25.5/23.5/53.8 %	0.0/0.0/75.0 %	13.8/30.8/78.8 %	28.0/12.7/58.8 %	36.4/32.0/80.6 %	0.0/0.0/91.9 %	17.3/16.5/73.1 %	19.4 %
SURF/$\widehat{\text{EMD}}$	6.2 %	40.4/33.3/58.8 %	16.7/22.7/73.8 %	10.3/16.7/74.4 %	16.0/15.4/73.1 %	4.5/4.2/72.5 %	0.0/0.0/87.5 %	14.7/15.4/73.3 %	20.0 %
Majority vote	22.5 %	42.6/47.6/69.4 %	10.0/21.4/76.2 %	34.5/52.6/82.5 %	56.0/23.3/64.4 %	27.3/42.9/85.0 %	14.3/9.1/90.0 %	30.8/32.8/77.9 %	33.8 %

E.6 Canny/Chamfer/None (45°)

Feature/Metric	Acc	G0	G1	G2	G3	G5	G6	Avg	TGA
Dummy	26.2 %	100.0/29.4/29.4 %	0.0/0.0/81.2 %	0.0/0.0/81.9 %	0.0/0.0/84.4 %	0.0/0.0/86.2 %	0.0/0.0/95.6 %	16.7/4.9/76.5 %	29.4 %
Contour/Cf	15.0 %	80.9/29.5/37.5 %	20.0/54.5/81.9 %	3.4/25.0/80.6 %	12.0/20.0/78.8 %	0.0/0.0/86.2 %	0.0/0.0/95.0 %	19.4/21.5/76.7 %	30.0 %
Canny/Cf	8.1 %	80.9/27.1/30.6 %	0.0/0.0/80.0 %	0.0/0.0/81.9 %	4.0/6.7/76.2 %	4.5/33.3/85.6 %	0.0/0.0/95.6 %	14.9/11.2/75.0 %	25.0 %
HOG/Eu	11.9 %	70.2/32.7/48.8 %	20.0/26.1/74.4 %	10.3/42.9/81.2 %	16.0/20.0/76.9 %	0.0/0.0/84.4 %	0.0/0.0/91.9 %	19.4/20.3/76.2 %	28.7 %
HOG/χ^2	11.2 %	68.1/32.0/48.1 %	16.7/19.2/71.2 %	6.9/40.0/81.2 %	16.0/22.2/78.1 %	0.0/0.0/84.4 %	0.0/0.0/90.6 %	17.9/18.9/75.6 %	26.9 %
PHOG/Eu	13.1 %	42.6/32.3/56.9 %	16.7/33.3/78.1 %	13.8/26.7/77.5 %	32.0/21.6/71.2 %	22.7/20.8/77.5 %	14.3/14.3/92.5 %	23.7/24.8/75.6 %	26.9 %
PHOG/χ^2	11.2 %	59.6/32.6/51.9 %	10.0/30.0/78.8 %	3.4/9.1/76.2 %	36.0/20.9/68.8 %	13.6/42.9/85.6 %	0.0/0.0/93.8 %	20.4/22.6/75.8 %	27.5 %
THOG/EMD	12.5 %	74.5/31.2/44.4 %	13.3/21.1/74.4 %	3.4/25.0/80.6 %	12.0/17.6/77.5 %	9.1/66.7/86.9 %	0.0/0.0/92.5 %	18.7/26.9/76.0 %	28.1 %
THOG/$\widehat{\text{EMD}}$	10.0 %	57.4/27.6/43.1 %	16.7/25.0/75.0 %	3.4/14.3/78.8 %	24.0/23.1/75.6 %	0.0/0.0/85.0 %	14.3/14.3/92.5 %	19.3/17.4/75.0 %	25.0 %
SIFT/EMD	8.8 %	10.6/22.7/63.1 %	0.0/0.0/81.2 %	6.9/13.3/75.0 %	64.0/16.7/44.4 %	18.2/16.0/75.6 %	0.0/0.0/94.4 %	16.6/11.5/72.3 %	16.9 %
SURF/EMD	8.1 %	14.9/41.2/68.8 %	13.3/22.2/75.0 %	13.8/30.8/78.8 %	56.0/17.5/51.9 %	4.5/3.3/68.8 %	0.0/0.0/94.4 %	17.1/19.2/72.9 %	18.8 %
SIFT/$\widehat{\text{EMD}}$	10.6 %	25.5/37.5/65.6 %	6.7/13.3/74.4 %	10.3/20.0/76.2 %	36.0/16.4/61.3 %	22.7/12.8/68.1 %	0.0/0.0/93.1 %	16.9/16.7/73.1 %	19.4 %
SURF/$\widehat{\text{EMD}}$	7.5 %	19.1/47.4/70.0 %	13.3/13.8/68.1 %	6.9/11.8/73.8 %	60.0/23.4/63.1 %	4.5/6.7/78.1 %	0.0/0.0/85.6 %	17.3/17.2/73.1 %	19.4 %
Majority vote	11.2 %	68.1/29.6/43.1 %	10.0/25.0/77.5 %	3.4/20.0/80.0 %	32.0/26.7/75.6 %	0.0/0.0/85.6 %	0.0/0.0/93.1 %	18.9/16.9/75.8 %	27.5 %

E.7 Canny/Chamfer/None (30°)

Feature/Metric	Acc	G0	G1	G2	G3	G5	G6	Avg	TGA
Dummy	26.2 %	100.0/29.4/29.4 %	0.0/0.0/81.2 %	0.0/0.0/81.9 %	0.0/0.0/84.4 %	0.0/0.0/86.2 %	0.0/0.0/95.6 %	16.7/4.9/76.5 %	29.4 %
Contour/Cf	13.8 %	74.5/29.7/40.6 %	16.7/38.5/79.4 %	0.0/0.0/80.6 %	16.0/17.4/75.0 %	4.5/50.0/86.2 %	0.0/0.0/94.4 %	18.6/22.6/76.0 %	28.1 %
Canny/Cf	5.6 %	78.7/27.2/31.9 %	3.3/33.3/80.6 %	0.0/0.0/81.9 %	8.0/10.5/75.0 %	0.0/0.0/85.6 %	0.0/0.0/95.0 %	15.0/11.8/75.0 %	25.0 %
HOG/Eu	12.5 %	61.7/27.4/40.6 %	26.7/28.6/73.8 %	3.4/25.0/80.6 %	8.0/11.1/75.6 %	0.0/0.0/85.6 %	0.0/0.0/93.8 %	16.6/15.3/75.0 %	25.0 %
HOG/χ^2	14.4 %	63.8/26.8/38.1 %	30.0/31.0/74.4 %	3.4/50.0/81.9 %	8.0/14.3/78.1 %	0.0/0.0/86.2 %	0.0/0.0/93.8 %	17.5/20.4/75.4 %	26.2 %
PHOG/Eu	11.2 %	48.9/34.8/58.1 %	13.3/30.8/78.1 %	10.3/15.8/73.8 %	32.0/19.0/68.1 %	4.5/5.0/75.0 %	0.0/0.0/95.6 %	18.2/17.6/74.8 %	24.4 %
PHOG/χ^2	13.1 %	61.7/33.3/52.5 %	10.0/30.0/78.8 %	6.9/16.7/76.9 %	36.0/21.4/69.4 %	4.5/11.1/81.9 %	0.0/0.0/95.6 %	19.9/18.8/75.8 %	27.5 %
THOG/EMD	13.8 %	57.4/28.1/44.4 %	20.0/26.1/74.4 %	3.4/12.5/78.1 %	16.0/13.8/71.2 %	0.0/0.0/84.4 %	0.0/0.0/95.0 %	16.1/13.4/74.6 %	23.8 %
THOG/$\widehat{\text{EMD}}$	11.2 %	40.4/27.5/51.2 %	23.3/21.9/70.0 %	0.0/0.0/79.4 %	24.0/15.0/66.9 %	0.0/0.0/83.8 %	28.6/18.2/91.2 %	19.4/13.8/73.8 %	21.2 %
SIFT/EMD	7.5 %	10.6/26.3/65.0 %	0.0/0.0/80.6 %	6.9/13.3/75.0 %	48.0/17.6/56.9 %	45.5/18.2/64.4 %	0.0/0.0/94.4 %	18.5/12.6/72.7 %	18.1 %
SURF/EMD	8.1 %	12.8/19.4/58.8 %	0.0/0.0/79.4 %	17.2/22.7/74.4 %	44.0/14.1/49.4 %	4.5/4.3/73.1 %	0.0/0.0/93.8 %	13.1/10.1/71.5 %	14.4 %
SIFT/$\widehat{\text{EMD}}$	7.5 %	21.3/37.0/66.2 %	6.7/15.4/75.6 %	6.9/14.3/75.6 %	48.0/20.0/61.9 %	36.4/19.0/70.0 %	0.0/0.0/93.1 %	19.9/17.6/73.8 %	21.2 %
SURF/$\widehat{\text{EMD}}$	10.0 %	12.8/26.1/63.7 %	50.0/21.7/56.9 %	17.2/20.8/73.1 %	8.0/8.7/72.5 %	4.5/14.3/83.1 %	0.0/0.0/86.9 %	15.4/15.3/72.7 %	18.1 %
Majority vote	13.8 %	59.6/27.7/42.5 %	23.3/33.3/76.9 %	3.4/16.7/79.4 %	16.0/13.8/71.2 %	0.0/0.0/85.0 %	0.0/0.0/95.0 %	17.1/15.3/75.0 %	25.0 %

E.8 Canny/Chamfer/None (22.5°)

Feature/Metric	Acc	G0	G1	G2	G3	G5	G6	Avg	TGA
Dummy	26.2 %	100.0/29.4/29.4 %	0.0/0.0/81.2 %	0.0/0.0/81.9 %	0.0/0.0/84.4 %	0.0/0.0/86.2 %	0.0/0.0/95.6 %	16.7/4.9/76.5 %	29.4 %
Contour/Cf	10.6 %	74.5/28.5/37.5 %	13.3/33.3/78.8 %	0.0/0.0/80.6 %	28.0/31.8/79.4 %	0.0/0.0/85.6 %	0.0/0.0/95.6 %	19.3/15.6/76.2 %	28.7 %
Canny/Cf	6.2 %	85.1/27.8/30.6 %	3.3/100.0/81.9 %	0.0/0.0/81.2 %	4.0/8.3/78.1 %	4.5/50.0/86.2 %	0.0/0.0/95.6 %	16.2/31.0/75.6 %	26.9 %
HOG/Eu	12.5 %	63.8/28.8/43.1 %	26.7/30.8/75.0 %	0.0/0.0/81.2 %	8.0/11.8/76.2 %	4.5/25.0/85.0 %	0.0/0.0/90.6 %	17.2/16.1/75.2 %	25.6 %
HOG/χ^2	15.6 %	68.1/28.8/41.2 %	26.7/32.0/75.6 %	0.0/0.0/80.6 %	4.0/6.2/75.6 %	4.5/25.0/85.0 %	0.0/0.0/94.4 %	17.2/15.3/75.4 %	26.2 %
PHOG/Eu	11.9 %	53.2/31.2/51.9 %	0.0/0.0/76.9 %	3.4/14.3/78.8 %	28.0/13.2/60.0 %	9.1/16.7/81.2 %	0.0/0.0/95.0 %	15.6/12.6/74.0 %	21.9 %
PHOG/χ^2	13.8 %	61.7/33.3/52.5 %	0.0/0.0/79.4 %	3.4/25.0/80.6 %	40.0/17.2/60.6 %	9.1/25.0/83.8 %	0.0/0.0/95.6 %	19.0/16.8/75.4 %	26.2 %
THOG/EMD	13.8 %	74.5/34.7/51.2 %	13.3/22.2/75.0 %	3.4/50.0/81.9 %	28.0/19.4/70.6 %	4.5/50.0/86.2 %	0.0/0.0/95.0 %	20.6/29.4/76.7 %	30.0 %
THOG/$\widehat{\text{EMD}}$	12.5 %	59.6/31.1/49.4 %	33.3/27.8/71.2 %	3.4/33.3/81.2 %	20.0/29.4/80.0 %	0.0/0.0/83.8 %	28.6/20.0/91.9 %	24.2/23.6/76.2 %	28.7 %
SIFT/EMD	10.0 %	19.1/42.9/68.8 %	3.3/14.3/78.1 %	13.8/25.0/76.9 %	16.0/7.5/56.2 %	18.2/7.0/55.6 %	0.0/0.0/91.9 %	11.7/16.1/71.2 %	13.8 %
SURF/EMD	3.8 %	17.0/28.6/63.1 %	6.7/12.5/73.8 %	3.4/12.5/78.1 %	28.0/9.7/48.1 %	13.6/13.0/75.6 %	0.0/0.0/87.5 %	11.5/12.7/71.0 %	13.1 %
SIFT/$\widehat{\text{EMD}}$	12.5 %	25.5/41.4/67.5 %	10.0/16.7/73.8 %	6.9/13.3/75.0 %	12.0/7.7/63.7 %	45.5/22.2/70.6 %	0.0/0.0/86.9 %	16.6/16.9/72.9 %	18.8 %
SURF/$\widehat{\text{EMD}}$	10.0 %	23.4/34.4/64.4 %	26.7/19.0/65.0 %	20.7/28.6/76.2 %	16.0/10.3/65.0 %	9.1/33.3/85.0 %	0.0/0.0/83.1 %	16.0/20.9/73.1 %	19.4 %
Majority vote	15.0 %	72.3/31.2/45.0 %	20.0/35.3/78.1 %	3.4/50.0/81.9 %	16.0/15.4/73.1 %	9.1/40.0/85.6 %	0.0/0.0/95.0 %	20.1/28.6/76.5 %	29.4 %

E.9 Canny/Chamfer/Gradient descent

Feature/Metric	Acc	G0	G1	G2	G3	G5	G6	Avg	TGA
Dummy	26.2 %	100.0/29.4/29.4 %	0.0/0.0/81.2 %	0.0/0.0/81.9 %	0.0/0.0/84.4 %	0.0/0.0/86.2 %	0.0/0.0/95.6 %	16.7/4.9/76.5 %	29.4 %
Contour/Cf	20.0 %	57.4/46.6/68.1 %	13.3/25.0/76.2 %	27.6/47.1/81.2 %	36.0/21.4/69.4 %	22.7/25.0/80.0 %	0.0/0.0/91.2 %	26.2/27.5/77.7 %	33.1 %
Canny/Cf	21.2 %	66.0/37.8/58.1 %	3.3/16.7/78.8 %	20.7/66.7/83.8 %	40.0/25.0/71.9 %	27.3/27.3/80.0 %	0.0/0.0/95.0 %	26.2/28.9/77.9 %	33.8 %
HOG/Eu	18.8 %	46.8/53.7/72.5 %	20.0/23.1/72.5 %	27.6/38.1/78.8 %	24.0/14.6/66.2 %	31.8/35.0/82.5 %	14.3/9.1/90.0 %	27.4/28.9/77.1 %	31.2 %
HOG/χ^2	18.8 %	48.9/53.5/72.5 %	23.3/29.2/75.0 %	31.0/39.1/78.8 %	28.0/17.1/67.5 %	27.3/31.6/81.9 %	0.0/0.0/89.4 %	26.4/28.4/77.5 %	32.5 %
PHOG/Eu	13.8 %	34.0/43.2/67.5 %	30.0/30.0/73.8 %	13.8/21.1/75.0 %	36.0/19.1/66.2 %	4.5/10.0/81.2 %	0.0/0.0/85.0 %	19.7/20.6/74.8 %	24.4 %
PHOG/χ^2	14.4 %	44.7/42.0/65.6 %	26.7/36.4/77.5 %	10.3/16.7/74.4 %	36.0/22.0/70.0 %	13.6/20.0/80.6 %	0.0/0.0/86.9 %	21.9/22.8/75.8 %	27.5 %
THOG/EMD	16.2 %	51.1/50.0/70.6 %	20.0/40.0/79.4 %	17.2/38.5/80.0 %	56.0/28.6/71.2 %	31.8/29.2/80.0 %	14.3/9.1/90.0 %	31.7/32.5/78.5 %	35.6 %
THOG/$\widehat{\text{EMD}}$	7.5 %	29.8/25.9/54.4 %	23.3/31.8/76.2 %	0.0/0.0/78.1 %	20.0/15.2/70.0 %	18.2/11.8/70.0 %	14.3/9.1/90.0 %	17.6/15.6/73.1 %	19.4 %
SIFT/EMD	7.5 %	23.4/23.4/55.0 %	3.3/14.3/78.1 %	6.9/20.0/78.1 %	40.0/16.7/59.4 %	13.6/10.3/71.9 %	0.0/0.0/91.2 %	14.5/14.1/72.3 %	16.9 %
SURF/EMD	7.5 %	27.7/35.1/63.7 %	0.0/0.0/76.2 %	10.3/21.4/76.9 %	56.0/18.2/53.8 %	13.6/16.7/78.8 %	0.0/0.0/91.9 %	17.9/15.2/73.5 %	20.6 %
SIFT/$\widehat{\text{EMD}}$	11.2 %	31.9/33.3/61.3 %	10.0/27.3/78.1 %	17.2/31.2/78.1 %	36.0/17.3/63.1 %	13.6/12.0/74.4 %	0.0/0.0/88.8 %	18.1/20.2/74.0 %	21.9 %
SURF/$\widehat{\text{EMD}}$	7.5 %	14.9/18.4/55.6 %	10.0/10.0/66.2 %	13.8/25.0/76.9 %	20.0/18.5/73.8 %	18.2/10.3/66.9 %	0.0/0.0/89.4 %	12.8/13.7/71.5 %	14.4 %
Majority vote	19.4 %	48.9/51.1/71.2 %	23.3/38.9/78.8 %	24.1/41.2/80.0 %	44.0/21.2/65.6 %	27.3/31.6/81.9 %	0.0/0.0/90.0 %	27.9/30.7/77.9 %	33.8 %

E.10 HOG/Euclidean/None (45°)

Feature/Metric	Acc	G0	G1	G2	G3	G5	G6	Avg	TGA
Dummy	26.2 %	100.0/29.4/29.4 %	0.0/0.0/81.2 %	0.0/0.0/81.9 %	0.0/0.0/84.4 %	0.0/0.0/86.2 %	0.0/0.0/95.6 %	16.7/4.9/76.5 %	29.4 %
Contour/Cf	12.5 %	78.7/28.7/36.2 %	13.3/26.7/76.9 %	0.0/0.0/80.6 %	4.0/10.0/79.4 %	0.0/0.0/85.6 %	0.0/0.0/93.8 %	16.0/10.9/75.4 %	26.2 %
Canny/Cf	11.2 %	85.1/28.6/33.1 %	0.0/0.0/78.8 %	0.0/0.0/81.9 %	8.0/16.7/79.4 %	4.5/33.3/85.6 %	0.0/0.0/95.0 %	16.3/13.1/75.6 %	26.9 %
HOG/Eu	8.8 %	51.1/26.7/44.4 %	30.0/36.0/76.9 %	0.0/0.0/78.1 %	12.0/13.6/74.4 %	0.0/0.0/77.5 %	0.0/0.0/93.8 %	15.5/12.7/74.2 %	22.5 %
HOG/χ^2	8.8 %	53.2/25.8/41.2 %	23.3/26.9/73.8 %	0.0/0.0/80.0 %	12.0/15.0/75.6 %	0.0/0.0/80.0 %	0.0/0.0/93.1 %	14.8/11.3/74.0 %	21.9 %
PHOG/Eu	15.0 %	53.2/29.1/48.1 %	6.7/22.2/78.1 %	6.9/14.3/75.6 %	32.0/19.0/68.1 %	4.5/16.7/83.8 %	0.0/0.0/93.8 %	17.2/16.9/74.6 %	23.8 %
PHOG/χ^2	16.9 %	72.3/33.3/49.4 %	10.0/30.0/78.8 %	6.9/25.0/79.4 %	24.0/17.6/70.6 %	4.5/25.0/85.0 %	0.0/0.0/94.4 %	19.6/21.8/76.2 %	28.7 %
THOG/EMD	13.8 %	76.6/29.8/40.0 %	6.7/14.3/75.0 %	3.4/20.0/80.0 %	8.0/15.4/78.8 %	0.0/0.0/84.4 %	0.0/0.0/93.1 %	15.8/13.2/75.2 %	25.6 %
THOG/$\widehat{\text{EMD}}$	10.0 %	61.7/32.2/50.6 %	10.0/10.0/66.2 %	6.9/28.6/80.0 %	28.0/25.9/76.2 %	4.5/50.0/86.2 %	0.0/0.0/93.1 %	18.5/24.5/75.4 %	26.2 %
SIFT/EMD	10.0 %	14.9/36.8/67.5 %	0.0/0.0/81.2 %	0.0/0.0/78.1 %	52.0/14.0/42.5 %	40.9/25.0/75.0 %	0.0/0.0/91.9 %	18.0/12.6/72.7 %	18.1 %
SURF/EMD	5.6 %	27.7/39.4/66.2 %	6.7/50.0/81.2 %	0.0/0.0/80.6 %	72.0/18.8/46.9 %	4.5/4.2/72.5 %	0.0/0.0/95.0 %	18.5/18.7/73.8 %	21.2 %
SIFT/$\widehat{\text{EMD}}$	11.2 %	23.4/42.3/68.1 %	0.0/0.0/76.9 %	20.7/31.6/77.5 %	40.0/16.7/59.4 %	54.5/28.6/75.0 %	0.0/0.0/91.9 %	23.1/19.9/74.8 %	24.4 %
SURF/$\widehat{\text{EMD}}$	7.5 %	21.3/33.3/64.4 %	3.3/14.3/78.1 %	20.7/16.2/66.2 %	44.0/20.8/65.0 %	9.1/10.0/76.2 %	0.0/0.0/87.5 %	16.4/15.8/72.9 %	18.8 %
Majority vote	11.9 %	72.3/30.1/42.5 %	10.0/30.0/78.8 %	0.0/0.0/80.6 %	20.0/16.7/71.9 %	0.0/0.0/85.0 %	0.0/0.0/93.8 %	17.1/12.8/75.4 %	26.2 %

E.11 HOG/Euclidean/None (30°)

Feature/Metric	Acc	G0	G1	G2	G3	G5	G6	Avg	TGA
Dummy	26.2 %	100.0/29.4/29.4 %	0.0/0.0/81.2 %	0.0/0.0/81.9 %	0.0/0.0/84.4 %	0.0/0.0/86.2 %	0.0/0.0/95.6 %	16.7/4.9/76.5 %	29.4 %
Contour/Cf	11.9 %	74.5/28.0/36.2 %	13.3/33.3/78.8 %	0.0/0.0/80.0 %	8.0/15.4/78.8 %	0.0/0.0/84.4 %	0.0/0.0/93.1 %	16.0/12.8/75.2 %	25.6 %
Canny/Cf	9.4 %	76.6/26.9/31.9 %	3.3/16.7/78.8 %	0.0/0.0/81.2 %	8.0/15.4/78.8 %	0.0/0.0/83.8 %	0.0/0.0/94.4 %	14.7/9.8/74.8 %	24.4 %
HOG/Eu	10.0 %	38.3/18.9/33.8 %	43.3/38.2/76.2 %	6.9/50.0/81.9 %	4.0/12.5/80.6 %	0.0/0.0/76.2 %	0.0/0.0/93.8 %	15.4/19.9/73.8 %	21.2 %
HOG/χ^2	12.5 %	46.8/21.8/35.0 %	43.3/36.1/75.0 %	0.0/0.0/80.6 %	12.0/37.5/83.1 %	0.0/0.0/79.4 %	0.0/0.0/94.4 %	17.0/15.9/74.6 %	23.8 %
PHOG/Eu	16.2 %	57.4/45.0/66.9 %	3.3/25.0/80.0 %	10.3/20.0/76.2 %	56.0/20.6/59.4 %	4.5/8.3/80.0 %	0.0/0.0/95.0 %	21.9/19.8/76.2 %	28.7 %
PHOG/χ^2	15.0 %	72.3/40.0/60.0 %	6.7/50.0/81.2 %	3.4/14.3/78.8 %	44.0/19.3/62.5 %	9.1/28.6/84.4 %	0.0/0.0/95.6 %	22.6/25.4/77.1 %	31.2 %
THOG/EMD	16.9 %	57.4/29.0/46.2 %	26.7/25.8/71.9 %	3.4/12.5/78.1 %	20.0/23.8/77.5 %	0.0/0.0/85.6 %	0.0/0.0/91.9 %	17.9/15.2/75.2 %	25.6 %
THOG/$\widehat{\text{EMD}}$	11.2 %	68.1/39.0/59.4 %	26.7/28.6/73.8 %	3.4/16.7/79.4 %	20.0/14.3/68.8 %	0.0/0.0/83.8 %	0.0/0.0/92.5 %	19.7/16.4/76.2 %	28.7 %
SIFT/EMD	3.1 %	14.9/31.8/65.6 %	0.0/0.0/80.6 %	3.4/7.7/75.0 %	32.0/10.1/45.0 %	18.2/9.5/65.0 %	0.0/0.0/93.8 %	11.4/9.9/70.8 %	12.5 %
SURF/EMD	11.9 %	34.0/34.0/61.3 %	3.3/14.3/78.1 %	10.3/13.6/71.9 %	32.0/13.3/56.9 %	9.1/9.1/75.0 %	0.0/0.0/94.4 %	14.8/14.1/72.9 %	18.8 %
SIFT/$\widehat{\text{EMD}}$	11.2 %	31.9/35.7/63.1 %	0.0/0.0/77.5 %	10.3/21.4/76.9 %	24.0/11.3/58.8 %	36.4/22.2/73.8 %	0.0/0.0/90.0 %	17.1/15.1/73.3 %	20.0 %
SURF/$\widehat{\text{EMD}}$	10.0 %	29.8/40.0/66.2 %	30.0/25.7/70.6 %	17.2/15.6/68.1 %	32.0/20.5/70.0 %	0.0/0.0/79.4 %	0.0/0.0/90.6 %	18.2/17.0/74.2 %	22.5 %
Majority vote	15.6 %	74.5/32.1/46.2 %	26.7/36.4/77.5 %	0.0/0.0/80.6 %	12.0/12.5/73.1 %	0.0/0.0/85.6 %	0.0/0.0/94.4 %	18.9/13.5/76.2 %	28.7 %

E.12 HOG/Euclidean/None (22.5°)

Feature/Metric	Acc	G0	G1	G2	G3	G5	G6	Avg	TGA
Dummy	26.2 %	100.0/29.4/29.4 %	0.0/0.0/81.2 %	0.0/0.0/81.9 %	0.0/0.0/84.4 %	0.0/0.0/86.2 %	0.0/0.0/95.6 %	16.7/4.9/76.5 %	29.4 %
Contour/Cf	13.1 %	74.5/28.9/38.8 %	6.7/22.2/78.1 %	0.0/0.0/80.0 %	20.0/20.8/75.6 %	4.5/33.3/85.6 %	0.0/0.0/95.6 %	17.6/17.6/75.6 %	26.9 %
Canny/Cf	8.1 %	87.2/30.8/38.8 %	3.3/25.0/80.0 %	0.0/0.0/80.6 %	16.0/22.2/78.1 %	4.5/100.0/86.9 %	0.0/0.0/94.4 %	18.5/29.7/76.5 %	29.4 %
HOG/Eu	11.2 %	38.3/18.0/30.6 %	30.0/33.3/75.6 %	3.4/33.3/81.2 %	8.0/22.2/81.2 %	4.5/5.3/75.6 %	0.0/0.0/94.4 %	14.0/18.7/73.1 %	19.4 %
HOG/χ^2	11.9 %	48.9/22.5/35.6 %	26.7/33.3/76.2 %	3.4/20.0/80.0 %	8.0/13.3/77.5 %	4.5/7.7/79.4 %	0.0/0.0/95.0 %	15.3/16.2/74.0 %	21.9 %
PHOG/Eu	13.8 %	55.3/32.1/52.5 %	6.7/25.0/78.8 %	6.9/40.0/81.2 %	52.0/21.3/62.5 %	0.0/0.0/83.1 %	0.0/0.0/95.6 %	20.1/19.7/75.6 %	26.9 %
PHOG/χ^2	14.4 %	61.7/30.5/47.5 %	6.7/50.0/81.2 %	0.0/0.0/80.6 %	48.0/21.8/65.0 %	0.0/0.0/84.4 %	0.0/0.0/95.0 %	19.4/17.1/75.6 %	26.9 %
THOG/EMD	14.4 %	78.7/34.3/49.4 %	13.3/18.2/72.5 %	0.0/0.0/80.0 %	24.0/26.1/77.5 %	4.5/25.0/85.0 %	0.0/0.0/95.6 %	20.1/17.3/76.7 %	30.0 %
THOG/$\widehat{\text{EMD}}$	11.2 %	44.7/28.0/50.0 %	33.3/25.0/68.8 %	0.0/0.0/80.0 %	40.0/35.7/79.4 %	0.0/0.0/83.1 %	0.0/0.0/90.0 %	19.7/14.8/75.2 %	25.6 %
SIFT/EMD	6.2 %	10.6/23.8/63.7 %	0.0/0.0/79.4 %	10.3/21.4/76.9 %	36.0/12.5/50.6 %	22.7/10.9/63.7 %	0.0/0.0/93.1 %	13.3/11.4/71.2 %	13.8 %
SURF/EMD	9.4 %	27.7/40.6/66.9 %	3.3/11.1/76.9 %	13.8/36.4/80.0 %	40.0/11.8/43.8 %	4.5/4.5/73.8 %	0.0/0.0/95.0 %	14.9/17.4/72.7 %	18.1 %
SIFT/$\widehat{\text{EMD}}$	8.8 %	17.0/22.2/58.1 %	10.0/21.4/76.2 %	6.9/18.2/77.5 %	32.0/17.8/66.2 %	27.3/12.8/64.4 %	0.0/0.0/91.2 %	15.5/15.4/72.3 %	16.9 %
SURF/$\widehat{\text{EMD}}$	7.5 %	27.7/29.5/59.4 %	26.7/21.1/67.5 %	13.8/30.8/78.8 %	24.0/15.8/68.1 %	4.5/11.1/81.9 %	14.3/5.6/85.6 %	18.5/19.0/73.5 %	20.6 %
Majority vote	15.6 %	66.0/28.2/40.6 %	20.0/31.6/76.9 %	0.0/0.0/80.6 %	24.0/22.2/75.0 %	0.0/0.0/85.0 %	0.0/0.0/95.6 %	18.3/13.7/75.6 %	26.9 %

E.13 HOG/Euclidean/Gradient descent

Feature/Metric	Acc	G0	G1	G2	G3	G5	G6	Avg	TGA
Dummy	26.2 %	100.0/29.4/29.4 %	0.0/0.0/81.2 %	0.0/0.0/81.9 %	0.0/0.0/84.4 %	0.0/0.0/86.2 %	0.0/0.0/95.6 %	16.7/4.9/76.5 %	29.4 %
Contour/Cf	23.1 %	57.4/55.1/73.8 %	26.7/36.4/77.5 %	20.7/31.6/77.5 %	60.0/28.3/70.0 %	22.7/45.5/85.6 %	0.0/0.0/91.9 %	31.3/32.8/79.4 %	38.1 %
Canny/Cf	20.6 %	57.4/42.9/65.0 %	20.0/40.0/79.4 %	13.8/33.3/79.4 %	48.0/24.5/68.8 %	27.3/28.6/80.6 %	0.0/0.0/95.6 %	27.8/28.2/78.1 %	34.4 %
HOG/Eu	26.2 %	61.7/54.7/73.8 %	26.7/26.7/72.5 %	20.7/60.0/83.1 %	64.0/33.3/74.4 %	18.2/44.4/85.6 %	0.0/0.0/89.4 %	31.9/36.5/79.8 %	39.4 %
HOG/χ^2	27.5 %	59.6/50.0/70.6 %	36.7/36.7/76.2 %	20.7/54.5/82.5 %	60.0/34.1/75.6 %	13.6/33.3/84.4 %	0.0/0.0/89.4 %	31.8/34.8/79.8 %	39.4 %
PHOG/Eu	10.0 %	29.8/32.6/61.3 %	20.0/31.6/76.9 %	13.8/26.7/77.5 %	40.0/15.2/55.6 %	4.5/10.0/81.2 %	0.0/0.0/91.2 %	18.0/19.3/74.0 %	21.9 %
PHOG/χ^2	15.0 %	53.2/41.0/63.7 %	10.0/33.3/79.4 %	10.3/17.6/75.0 %	44.0/19.3/62.5 %	4.5/10.0/81.2 %	0.0/0.0/91.9 %	20.3/20.2/75.6 %	26.9 %
THOG/EMD	21.9 %	51.1/52.2/71.9 %	20.0/31.6/76.9 %	13.8/26.7/77.5 %	56.0/24.1/65.6 %	13.6/21.4/81.2 %	14.3/12.5/91.9 %	28.1/28.1/77.5 %	32.5 %
THOG/$\widehat{\text{EMD}}$	19.4 %	48.9/44.2/66.9 %	13.3/21.1/74.4 %	10.3/25.0/78.1 %	44.0/23.9/69.4 %	22.7/26.3/80.6 %	0.0/0.0/88.1 %	23.2/23.4/76.2 %	28.7 %
SIFT/EMD	8.1 %	23.4/24.4/56.2 %	6.7/18.2/76.9 %	6.9/15.4/76.2 %	20.0/10.2/60.0 %	31.8/17.9/70.6 %	14.3/33.3/95.0 %	17.2/19.9/72.5 %	17.5 %
SURF/EMD	8.1 %	29.8/24.1/51.9 %	6.7/28.6/79.4 %	3.4/5.0/70.6 %	32.0/14.5/60.0 %	4.5/5.9/76.9 %	0.0/0.0/93.8 %	12.7/13.0/72.1 %	16.2 %
SIFT/$\widehat{\text{EMD}}$	11.2 %	29.8/32.6/61.3 %	6.7/16.7/76.2 %	13.8/22.2/75.6 %	20.0/12.2/65.0 %	22.7/12.2/66.9 %	14.3/20.0/93.8 %	17.9/19.3/73.1 %	19.4 %
SURF/$\widehat{\text{EMD}}$	7.5 %	29.8/28.6/57.5 %	13.3/21.1/74.4 %	6.9/9.1/70.6 %	16.0/15.4/73.1 %	9.1/5.9/67.5 %	14.3/10.0/90.6 %	14.9/15.0/72.3 %	16.9 %
Majority vote	22.5 %	51.1/48.0/69.4 %	16.7/33.3/78.1 %	17.2/38.5/80.0 %	64.0/24.6/63.7 %	4.5/12.5/82.5 %	0.0/0.0/90.0 %	25.6/26.2/77.3 %	31.9 %

E.14 HOG/Euclidean/L-BFGS

Feature/Metric	Acc	G0	G1	G2	G3	G5	G6	Avg	TGA
Dummy	26.2 %	100.0/29.4/29.4 %	0.0/0.0/81.2 %	0.0/0.0/81.9 %	0.0/0.0/84.4 %	0.0/0.0/86.2 %	0.0/0.0/95.6 %	16.7/4.9/76.5 %	29.4 %
Contour/Cf	11.9 %	66.0/26.1/35.0 %	3.3/20.0/79.4 %	6.9/11.8/73.8 %	8.0/13.3/77.5 %	0.0/0.0/84.4 %	0.0/0.0/95.0 %	14.0/11.9/74.2 %	22.5 %
Canny/Cf	10.6 %	63.8/24.0/30.0 %	3.3/16.7/78.8 %	3.4/10.0/76.9 %	8.0/14.3/78.1 %	0.0/0.0/83.8 %	0.0/0.0/95.0 %	13.1/10.8/73.8 %	21.2 %
HOG/Eu	8.8 %	66.0/26.7/36.9 %	0.0/0.0/78.1 %	0.0/0.0/74.4 %	8.0/7.7/70.6 %	0.0/0.0/85.6 %	0.0/0.0/95.6 %	12.3/5.7/73.5 %	20.6 %
HOG/χ^2	10.6 %	61.7/24.6/33.1 %	3.3/14.3/78.1 %	3.4/7.7/75.0 %	4.0/5.0/73.1 %	0.0/0.0/85.0 %	0.0/0.0/95.6 %	12.1/8.6/73.3 %	20.0 %
PHOG/Eu	13.1 %	61.7/31.9/50.0 %	0.0/0.0/81.2 %	17.2/13.5/65.0 %	24.0/21.4/74.4 %	4.5/25.0/85.0 %	0.0/0.0/95.6 %	17.9/15.3/75.2 %	25.6 %
PHOG/χ^2	13.8 %	63.8/30.0/45.6 %	0.0/0.0/80.0 %	10.3/10.0/66.9 %	24.0/24.0/76.2 %	0.0/0.0/84.4 %	0.0/0.0/95.6 %	16.4/10.7/74.8 %	24.4 %
THOG/EMD	11.2 %	61.7/27.4/40.6 %	0.0/0.0/77.5 %	3.4/5.0/70.6 %	8.0/7.7/70.6 %	4.5/50.0/86.2 %	0.0/0.0/95.6 %	12.9/15.0/73.5 %	20.6 %
THOG/$\widehat{\text{EMD}}$	11.9 %	66.0/30.7/46.2 %	6.7/33.3/80.0 %	10.3/11.1/68.8 %	16.0/21.1/77.5 %	0.0/0.0/82.5 %	0.0/0.0/95.0 %	16.5/16.0/75.0 %	25.0 %
SIFT/EMD	11.2 %	36.2/23.0/45.6 %	0.0/0.0/79.4 %	10.3/8.6/63.7 %	20.0/13.5/67.5 %	9.1/18.2/81.9 %	0.0/0.0/95.6 %	12.6/10.5/72.3 %	16.9 %
SURF/EMD	13.1 %	42.6/31.2/55.6 %	0.0/0.0/78.8 %	37.9/23.9/66.9 %	20.0/17.2/72.5 %	9.1/13.3/79.4 %	0.0/0.0/94.4 %	18.3/14.3/74.6 %	23.8 %
SIFT/$\widehat{\text{EMD}}$	13.1 %	57.4/30.3/48.8 %	0.0/0.0/78.8 %	10.3/11.1/68.8 %	8.0/8.3/71.9 %	0.0/0.0/76.2 %	0.0/0.0/95.6 %	12.6/8.3/73.3 %	20.0 %
SURF/$\widehat{\text{EMD}}$	9.4 %	42.6/28.6/51.9 %	0.0/0.0/78.1 %	34.5/16.9/57.5 %	12.0/17.6/77.5 %	4.5/11.1/81.9 %	0.0/0.0/95.6 %	15.6/12.4/73.8 %	21.2 %
Majority vote	11.9 %	59.6/26.4/39.4 %	3.3/25.0/80.0 %	10.3/13.0/71.2 %	12.0/12.5/73.1 %	0.0/0.0/84.4 %	0.0/0.0/95.6 %	14.2/12.8/74.0 %	21.9 %

E.15 Trimmed HOG (size 50)/$\widehat{\text{EMD}}$/None (45°)

Feature/Metric	Acc	G0	G1	G2	G3	G5	G6	Avg	TGA
Dummy	27.1 %	100.0/29.7/29.7 %	0.0/0.0/81.3 %	0.0/0.0/81.3 %	0.0/0.0/83.9 %	0.0/0.0/87.1 %	0.0/0.0/96.1 %	16.7/4.9/76.6 %	29.7 %
Contour/Cf	11.6 %	69.6/27.6/36.8 %	3.4/7.7/74.2 %	3.4/16.7/78.7 %	8.0/11.8/75.5 %	0.0/0.0/85.2 %	0.0/0.0/96.1 %	14.1/10.6/74.4 %	23.2 %
Canny/Cf	12.3 %	73.9/27.9/35.5 %	10.3/27.3/78.1 %	0.0/0.0/80.6 %	16.0/21.1/76.8 %	0.0/0.0/85.8 %	0.0/0.0/96.1 %	16.7/12.7/75.5 %	26.5 %
HOG/Eu	11.0 %	54.3/27.2/43.2 %	17.2/19.2/71.0 %	0.0/0.0/78.1 %	4.0/3.8/68.4 %	0.0/0.0/86.5 %	0.0/0.0/92.9 %	12.6/8.4/73.3 %	20.0 %
HOG/χ^2	11.0 %	54.3/27.5/43.9 %	13.8/16.0/70.3 %	0.0/0.0/77.4 %	4.0/3.8/68.4 %	0.0/0.0/86.5 %	0.0/0.0/92.3 %	12.0/7.9/73.1 %	19.4 %
PHOG/Eu	11.0 %	52.2/36.4/58.7 %	3.4/50.0/81.3 %	10.3/17.6/74.2 %	44.0/26.2/71.0 %	10.0/8.3/74.2 %	0.0/0.0/93.5 %	20.0/23.1/75.5 %	26.5 %
PHOG/χ^2	10.3 %	52.2/31.2/51.6 %	3.4/25.0/80.0 %	6.9/11.8/72.9 %	40.0/22.7/68.4 %	0.0/0.0/82.6 %	0.0/0.0/92.3 %	17.1/15.1/74.6 %	23.9 %
THOG/EMD	9.0 %	60.9/26.9/39.4 %	0.0/0.0/76.1 %	3.4/20.0/79.4 %	8.0/7.1/68.4 %	10.0/33.3/85.8 %	0.0/0.0/93.5 %	13.7/14.6/73.8 %	21.3 %
THOG/$\widehat{\text{EMD}}$	11.6 %	71.7/37.5/56.1 %	17.2/20.8/72.3 %	3.4/7.1/73.5 %	20.0/20.8/74.8 %	0.0/0.0/86.5 %	0.0/0.0/93.5 %	18.7/14.4/76.1 %	28.4 %
SIFT/EMD	7.1 %	6.5/18.8/63.9 %	0.0/0.0/81.3 %	3.4/14.3/78.1 %	44.0/15.7/52.9 %	45.0/17.6/65.8 %	0.0/0.0/89.0 %	16.5/11.1/71.8 %	15.5 %
SURF/EMD	1.9 %	26.1/31.6/61.3 %	3.4/16.7/78.7 %	0.0/0.0/79.4 %	68.0/18.3/45.8 %	5.0/20.0/85.2 %	0.0/0.0/89.7 %	17.1/14.4/73.3 %	20.0 %
SIFT/$\widehat{\text{EMD}}$	8.4 %	15.2/25.0/61.3 %	6.9/25.0/78.7 %	10.3/15.8/72.9 %	24.0/12.8/61.3 %	35.0/15.6/67.1 %	0.0/0.0/91.0 %	15.2/15.7/72.0 %	16.1 %
SURF/$\widehat{\text{EMD}}$	8.4 %	23.9/31.4/61.9 %	6.9/8.7/69.0 %	3.4/9.1/75.5 %	40.0/16.1/56.8 %	0.0/0.0/85.2 %	0.0/0.0/82.6 %	12.4/10.9/71.8 %	15.5 %
Majority vote	11.6 %	60.9/27.7/41.3 %	6.9/18.2/76.8 %	0.0/0.0/76.8 %	16.0/12.1/67.7 %	0.0/0.0/86.5 %	0.0/0.0/94.8 %	14.0/9.7/74.0 %	21.9 %

E.16 Trimmed HOG (size 50)/$\widehat{\text{EMD}}$/None ($30°$)

Feature/Metric	Acc	G0	G1	G2	G3	G5	G6	Avg	TGA
Dummy	27.1 %	100.0/29.7/29.7 %	0.0/0.0/81.3 %	0.0/0.0/81.3 %	0.0/0.0/83.9 %	0.0/0.0/87.1 %	0.0/0.0/96.1 %	16.7/4.9/76.6 %	29.7 %
Contour/Cf	12.3 %	63.0/30.5/46.5 %	10.3/20.0/75.5 %	3.4/14.3/78.1 %	16.0/13.8/70.3 %	5.0/16.7/84.5 %	0.0/0.0/94.2 %	16.3/15.9/74.8 %	24.5 %
Canny/Cf	9.0 %	71.7/32.0/46.5 %	3.4/14.3/78.1 %	6.9/66.7/81.9 %	24.0/15.8/67.1 %	0.0/0.0/85.2 %	0.0/0.0/95.5 %	17.7/21.5/75.7 %	27.1 %
HOG/Eu	12.9 %	60.9/30.1/46.5 %	17.2/27.8/76.1 %	6.9/40.0/80.6 %	16.0/14.3/71.0 %	5.0/50.0/87.1 %	0.0/0.0/90.3 %	17.7/27.0/75.3 %	25.8 %
HOG/χ^2	13.5 %	56.5/28.3/44.5 %	13.8/25.0/76.1 %	10.3/37.5/80.0 %	16.0/14.3/71.0 %	5.0/50.0/87.1 %	0.0/0.0/90.3 %	16.9/25.8/74.8 %	24.5 %
PHOG/Eu	8.4 %	34.8/34.0/60.6 %	10.3/27.3/78.1 %	10.3/18.8/74.8 %	32.0/13.1/54.8 %	0.0/0.0/78.7 %	0.0/0.0/91.6 %	14.6/15.5/73.1 %	19.4 %
PHOG/χ^2	10.3 %	39.1/30.5/55.5 %	6.9/28.6/79.4 %	10.3/20.0/75.5 %	32.0/12.5/52.9 %	0.0/0.0/84.5 %	0.0/0.0/92.3 %	14.7/15.3/73.3 %	20.0 %
THOG/EMD	11.0 %	43.5/30.3/53.5 %	13.8/19.0/72.9 %	6.9/16.7/76.1 %	20.0/13.5/66.5 %	5.0/11.1/82.6 %	0.0/0.0/89.7 %	14.9/15.1/73.5 %	20.6 %
THOG/$\widehat{\text{EMD}}$	10.3 %	60.9/35.0/54.8 %	10.3/14.3/71.6 %	3.4/16.7/78.7 %	32.0/21.1/69.7 %	5.0/16.7/84.5 %	0.0/0.0/93.5 %	18.6/17.3/75.5 %	26.5 %
SIFT/EMD	9.7 %	10.9/25.0/63.9 %	3.4/12.5/77.4 %	3.4/12.5/77.4 %	56.0/16.9/48.4 %	30.0/18.2/73.5 %	0.0/0.0/94.2 %	17.3/14.2/72.5 %	17.4 %
SURF/EMD	11.0 %	19.6/20.0/52.9 %	3.4/10.0/76.1 %	20.7/20.7/70.3 %	48.0/24.0/67.1 %	10.0/13.3/80.0 %	16.7/16.7/93.5 %	19.7/17.4/73.3 %	20.0 %
SIFT/$\widehat{\text{EMD}}$	12.9 %	34.8/41.0/65.8 %	3.4/11.1/76.8 %	3.4/14.3/78.1 %	28.0/13.5/59.4 %	45.0/20.9/71.0 %	0.0/0.0/92.9 %	19.1/16.8/74.0 %	21.9 %
SURF/$\widehat{\text{EMD}}$	7.7 %	34.8/28.1/54.2 %	13.8/17.4/71.6 %	13.8/19.0/72.9 %	12.0/8.1/63.9 %	5.0/33.3/86.5 %	16.7/7.1/88.4 %	16.0/18.8/72.9 %	18.7 %
Majority vote	12.3 %	47.8/28.6/49.0 %	10.3/23.1/76.8 %	10.3/37.5/80.0 %	28.0/15.9/64.5 %	0.0/0.0/84.5 %	0.0/0.0/90.3 %	16.1/17.5/74.2 %	22.6 %

E.17 Trimmed HOG (size 50)/$\widehat{\text{EMD}}$/None ($22.5°$)

Feature/Metric	Acc	G0	G1	G2	G3	G5	G6	Avg	TGA
Dummy	27.1 %	100.0/29.7/29.7 %	0.0/0.0/81.3 %	0.0/0.0/81.3 %	0.0/0.0/83.9 %	0.0/0.0/87.1 %	0.0/0.0/96.1 %	16.7/4.9/76.6 %	29.7 %
Contour/Cf	11.0 %	69.6/30.8/44.5 %	6.9/15.4/75.5 %	0.0/0.0/80.6 %	20.0/17.2/71.6 %	0.0/0.0/85.2 %	0.0/0.0/92.9 %	16.1/10.6/75.1 %	25.2 %
Canny/Cf	9.0 %	71.7/30.0/41.9 %	6.9/28.6/79.4 %	3.4/33.3/80.6 %	20.0/17.2/71.6 %	15.0/100.0/89.0 %	0.0/0.0/94.2 %	19.5/34.9/76.1 %	28.4 %
HOG/Eu	10.3 %	71.7/34.4/51.0 %	6.9/9.5/70.3 %	6.9/28.6/79.4 %	8.0/8.0/70.3 %	0.0/0.0/85.8 %	16.7/25.0/94.8 %	18.4/17.6/75.3 %	25.8 %
HOG/χ^2	10.3 %	69.6/34.0/51.0 %	10.3/11.1/67.7 %	3.4/33.3/80.6 %	12.0/11.1/70.3 %	0.0/0.0/86.5 %	0.0/0.0/94.2 %	15.9/14.9/75.1 %	25.2 %
PHOG/Eu	12.3 %	41.3/37.3/61.9 %	17.2/38.5/79.4 %	6.9/13.3/74.2 %	32.0/14.8/59.4 %	5.0/6.2/78.1 %	0.0/0.0/92.3 %	17.1/18.4/74.2 %	22.6 %
PHOG/χ^2	16.8 %	56.5/38.2/60.0 %	17.2/41.7/80.0 %	6.9/20.0/77.4 %	36.0/16.4/60.0 %	10.0/28.6/85.2 %	0.0/0.0/94.2 %	21.1/24.1/76.1 %	28.4 %
THOG/EMD	17.4 %	65.2/38.0/58.1 %	10.3/18.8/74.8 %	10.3/42.9/80.6 %	40.0/21.3/66.5 %	10.0/66.7/87.7 %	0.0/0.0/94.2 %	22.7/31.3/77.0 %	31.0 %
THOG/$\widehat{\text{EMD}}$	9.7 %	56.5/31.0/49.7 %	10.3/12.0/69.0 %	6.9/40.0/80.6 %	20.0/16.1/70.3 %	5.0/33.3/86.5 %	0.0/0.0/91.6 %	16.5/22.1/74.6 %	23.9 %
SIFT/EMD	10.3 %	19.6/25.7/59.4 %	3.4/14.3/78.1 %	6.9/12.5/73.5 %	44.0/20.0/62.6 %	20.0/12.1/71.0 %	0.0/0.0/90.3 %	15.7/14.1/72.5 %	17.4 %
SURF/EMD	3.9 %	26.1/29.3/59.4 %	6.9/16.7/76.1 %	0.0/0.0/79.4 %	60.0/18.1/49.7 %	0.0/0.0/82.6 %	0.0/0.0/90.3 %	15.5/10.7/72.9 %	18.7 %
SIFT/$\widehat{\text{EMD}}$	8.4 %	19.6/25.0/58.7 %	0.0/0.0/74.2 %	6.9/10.5/71.6 %	32.0/15.7/61.3 %	30.0/20.0/75.5 %	0.0/0.0/91.0 %	14.7/11.9/72.0 %	16.1 %
SURF/$\widehat{\text{EMD}}$	8.4 %	28.3/27.7/56.8 %	27.6/21.1/67.1 %	6.9/13.3/74.2 %	28.0/24.1/74.2 %	5.0/14.3/83.9 %	0.0/0.0/83.9 %	16.0/16.7/73.3 %	20.0 %
Majority vote	11.6 %	69.6/34.8/52.3 %	6.9/14.3/74.8 %	3.4/50.0/81.3 %	28.0/17.1/66.5 %	0.0/0.0/85.8 %	0.0/0.0/93.5 %	18.0/19.4/75.7 %	27.1 %

E.18 Trimmed HOG (size 50)/$\widehat{\text{EMD}}$/Gradient descent

Feature/Metric	Acc	G0	G1	G2	G3	G5	G6	Avg	TGA
Dummy	27.1 %	100.0/29.7/29.7 %	0.0/0.0/81.3 %	0.0/0.0/81.3 %	0.0/0.0/83.9 %	0.0/0.0/87.1 %	0.0/0.0/96.1 %	16.7/4.9/76.6 %	29.7 %
Contour/Cf	15.5 %	45.7/36.2/60.0 %	20.7/30.0/76.1 %	6.9/25.0/78.7 %	20.0/11.1/61.3 %	10.0/11.1/78.1 %	0.0/0.0/92.3 %	17.2/18.9/74.4 %	23.2 %
Canny/Cf	13.5 %	52.2/38.1/60.6 %	13.8/28.6/77.4 %	10.3/60.0/81.9 %	24.0/14.0/63.9 %	15.0/12.0/74.8 %	0.0/0.0/92.9 %	19.2/25.4/75.3 %	25.8 %
HOG/Eu	16.1 %	37.0/34.7/60.6 %	31.0/36.0/76.8 %	13.8/33.3/78.7 %	28.0/18.4/68.4 %	5.0/5.0/75.5 %	0.0/0.0/89.0 %	19.1/21.2/74.8 %	24.5 %
HOG/χ^2	15.5 %	37.0/34.7/60.6 %	27.6/36.4/77.4 %	13.8/33.3/78.7 %	32.0/18.2/65.8 %	5.0/5.9/77.4 %	0.0/0.0/89.0 %	19.2/21.4/74.8 %	24.5 %
PHOG/Eu	10.3 %	28.3/29.5/58.7 %	31.0/30.0/73.5 %	13.8/23.5/75.5 %	20.0/12.5/64.5 %	0.0/0.0/84.5 %	0.0/0.0/83.2 %	15.5/15.9/73.3 %	20.0 %
PHOG/χ^2	14.8 %	47.8/41.5/64.5 %	20.7/30.0/76.1 %	10.3/16.7/73.5 %	24.0/14.6/65.2 %	10.0/20.0/83.2 %	0.0/0.0/87.7 %	18.8/20.5/75.1 %	25.2 %
THOG/EMD	12.3 %	41.3/44.2/67.1 %	17.2/25.0/74.8 %	10.3/30.0/78.7 %	20.0/9.8/57.4 %	20.0/18.2/78.1 %	0.0/0.0/90.3 %	18.1/21.2/74.4 %	23.2 %
THOG/$\widehat{\text{EMD}}$	9.7 %	26.1/34.3/63.2 %	20.7/25.0/73.5 %	0.0/0.0/72.9 %	24.0/12.8/61.3 %	15.0/13.6/76.8 %	0.0/0.0/87.1 %	14.3/14.3/72.5 %	17.4 %
SIFT/EMD	5.2 %	17.4/20.0/54.8 %	3.4/10.0/76.1 %	6.9/22.2/78.1 %	28.0/12.1/55.5 %	20.0/14.3/74.2 %	16.7/10.0/91.0 %	15.4/14.8/71.6 %	14.8 %
SURF/EMD	8.4 %	21.7/30.3/61.9 %	17.2/33.3/78.1 %	6.9/13.3/74.2 %	36.0/16.1/59.4 %	5.0/4.2/72.9 %	16.7/8.3/89.7 %	17.3/17.6/72.7 %	18.1 %
SIFT/$\widehat{\text{EMD}}$	7.7 %	26.1/22.6/51.6 %	3.4/9.1/75.5 %	3.4/20.0/79.4 %	32.0/16.3/62.6 %	20.0/12.9/72.3 %	0.0/0.0/92.3 %	14.2/13.5/72.3 %	16.8 %
SURF/$\widehat{\text{EMD}}$	9.0 %	34.8/32.0/58.7 %	10.3/18.8/74.8 %	6.9/22.2/78.1 %	32.0/26.7/74.8 %	30.0/15.8/70.3 %	0.0/0.0/88.4 %	19.0/19.2/74.2 %	22.6 %
Majority vote	12.9 %	37.0/37.0/62.6 %	24.1/33.3/76.8 %	6.9/20.0/77.4 %	24.0/11.8/58.7 %	15.0/20.0/81.3 %	0.0/0.0/88.4 %	17.8/20.3/74.2 %	22.6 %

F Pairwise accuracies

Pairwise accuracies of different nearest-neighbor classifiers are presented below, analogously to Table 10. Each table corresponds to an individual fitting scheme.

F.1 Manually adjusted angles

Feature/Metric	all 12	1001/1201	1011/1031	1301/1521	1311/1341	1001/1130	1511/1521	1341/1511	1110/1601
Dummy	26.3 %	59.2 %	60.0 %	81.8 %	68.8 %	66.7 %	90.9 %	80.0 %	56.3 %
Contour/Cf	30.0 %	88.7 %	80.0 %	100.0 %	75.0 %	98.4 %	13.6 %	24.0 %	43.8 %
Canny/Cf	28.1 %	80.3 %	80.0 %	100.0 %	93.8 %	90.5 %	13.6 %	20.0 %	37.5 %
HOG/Eu	27.5 %	73.2 %	100.0 %	81.8 %	100.0 %	76.2 %	27.3 %	24.0 %	31.3 %
HOG/χ^2	26.9 %	74.6 %	100.0 %	72.7 %	100.0 %	79.4 %	31.8 %	20.0 %	37.5 %
PHOG/Eu	23.8 %	73.2 %	100.0 %	81.8 %	81.3 %	85.7 %	27.3 %	32.0 %	25.0 %
PHOG/χ^2	26.9 %	81.7 %	100.0 %	81.8 %	81.3 %	92.1 %	22.7 %	28.0 %	25.0 %
THOG/EMD	25.6 %	83.1 %	100.0 %	81.8 %	68.8 %	92.1 %	9.1 %	20.0 %	31.3 %
THOG/$\widehat{\text{EMD}}$	20.6 %	62.0 %	80.0 %	63.6 %	50.0 %	57.1 %	50.0 %	28.0 %	37.5 %
SIFT/EMD	6.3 %	52.1 %	20.0 %	36.4 %	37.5 %	50.8 %	54.5 %	56.0 %	43.8 %
SURF/EMD	8.8 %	57.7 %	80.0 %	45.5 %	68.8 %	65.1 %	27.3 %	36.0 %	31.3 %
SIFT/$\widehat{\text{EMD}}$	11.3 %	52.1 %	40.0 %	63.6 %	37.5 %	60.3 %	54.5 %	48.0 %	56.3 %
SURF/$\widehat{\text{EMD}}$	8.1 %	36.6 %	60.0 %	90.9 %	56.3 %	33.3 %	45.5 %	52.0 %	50.0 %
Majority vote	27.5 %	71.8 %	100.0 %	81.8 %	81.3 %	85.7 %	27.3 %	24.0 %	31.3 %

F.2 Contour/Chamfer/None (45°)

Feature/Metric	all 12	1001/1201	1011/1031	1301/1521	1311/1341	1001/1130	1511/1521	1341/1511	1110/1601
Dummy	26.3 %	59.2 %	60.0 %	81.8 %	68.8 %	66.7 %	90.9 %	80.0 %	56.3 %
Contour/Cf	13.8 %	59.2 %	40.0 %	54.5 %	56.3 %	52.4 %	9.1 %	20.0 %	37.5 %
Canny/Cf	10.0 %	54.9 %	40.0 %	45.5 %	68.8 %	57.1 %	9.1 %	20.0 %	18.8 %
HOG/Eu	9.4 %	50.7 %	40.0 %	72.7 %	62.5 %	34.9 %	40.9 %	40.0 %	37.5 %
HOG/χ^2	10.0 %	57.7 %	40.0 %	72.7 %	62.5 %	33.3 %	36.4 %	32.0 %	37.5 %
PHOG/Eu	14.4 %	62.0 %	40.0 %	63.6 %	81.3 %	65.1 %	45.5 %	44.0 %	62.5 %
PHOG/χ^2	14.4 %	63.4 %	40.0 %	72.7 %	68.8 %	61.9 %	31.8 %	36.0 %	43.8 %
THOG/EMD	11.9 %	57.7 %	60.0 %	45.5 %	75.0 %	58.7 %	27.3 %	36.0 %	31.3 %
THOG/$\widehat{\text{EMD}}$	8.1 %	39.4 %	60.0 %	63.6 %	43.8 %	33.3 %	22.7 %	16.0 %	25.0 %
SIFT/EMD	10.6 %	53.5 %	20.0 %	36.4 %	68.8 %	47.6 %	31.8 %	40.0 %	43.8 %
SURF/EMD	5.0 %	45.1 %	40.0 %	63.6 %	56.3 %	58.7 %	50.0 %	36.0 %	37.5 %
SIFT/$\widehat{\text{EMD}}$	12.5 %	66.2 %	20.0 %	18.2 %	25.0 %	60.3 %	59.1 %	48.0 %	50.0 %
SURF/$\widehat{\text{EMD}}$	8.8 %	45.1 %	60.0 %	63.6 %	56.3 %	33.3 %	40.9 %	36.0 %	31.3 %
Majority vote	12.5 %	62.0 %	60.0 %	72.7 %	62.5 %	47.6 %	27.3 %	24.0 %	43.8 %

F.3 Contour/Chamfer/None (30°)

Feature/Metric	all 12	1001/1201	1011/1031	1301/1521	1311/1341	1001/1130	1511/1521	1341/1511	1110/1601
Dummy	26.3 %	59.2 %	60.0 %	81.8 %	68.8 %	66.7 %	90.9 %	80.0 %	56.3 %
Contour/Cf	13.1 %	52.1 %	40.0 %	54.5 %	56.3 %	55.6 %	13.6 %	20.0 %	37.5 %
Canny/Cf	8.1 %	54.9 %	40.0 %	45.5 %	56.3 %	57.1 %	9.1 %	24.0 %	31.3 %
HOG/Eu	11.9 %	52.1 %	60.0 %	54.5 %	62.5 %	46.0 %	13.6 %	32.0 %	43.8 %
HOG/χ^2	15.0 %	52.1 %	40.0 %	81.8 %	62.5 %	52.4 %	13.6 %	32.0 %	31.3 %
PHOG/Eu	11.9 %	57.7 %	40.0 %	72.7 %	68.8 %	66.7 %	27.3 %	36.0 %	50.0 %
PHOG/χ^2	14.4 %	60.6 %	40.0 %	72.7 %	68.8 %	66.7 %	9.1 %	36.0 %	50.0 %
THOG/EMD	13.1 %	47.9 %	40.0 %	54.5 %	62.5 %	58.7 %	36.4 %	40.0 %	50.0 %
THOG/$\widehat{\text{EMD}}$	8.8 %	49.3 %	40.0 %	45.5 %	50.0 %	52.4 %	13.6 %	28.0 %	25.0 %
SIFT/EMD	11.3 %	52.1 %	60.0 %	54.5 %	50.0 %	55.6 %	31.8 %	36.0 %	50.0 %
SURF/EMD	17.5 %	46.5 %	60.0 %	18.2 %	56.3 %	61.9 %	50.0 %	32.0 %	50.0 %
SIFT/$\widehat{\text{EMD}}$	12.5 %	67.6 %	40.0 %	18.2 %	31.3 %	54.0 %	40.9 %	52.0 %	31.3 %
SURF/$\widehat{\text{EMD}}$	8.8 %	47.9 %	60.0 %	27.3 %	68.8 %	28.6 %	72.7 %	56.0 %	37.5 %
Majority vote	14.4 %	56.3 %	60.0 %	54.5 %	68.8 %	60.3 %	13.6 %	24.0 %	37.5 %

F.4 Contour/Chamfer/None (22.5°)

Feature/Metric	all 12	1001/1201	1011/1031	1301/1521	1311/1341	1001/1130	1511/1521	1341/1511	1110/1601
Dummy	26.3 %	59.2 %	60.0 %	81.8 %	68.8 %	66.7 %	90.9 %	80.0 %	56.3 %
Contour/Cf	14.4 %	50.7 %	40.0 %	54.5 %	62.5 %	54.0 %	13.6 %	24.0 %	37.5 %
Canny/Cf	8.8 %	54.9 %	40.0 %	36.4 %	56.3 %	60.3 %	13.6 %	24.0 %	37.5 %
HOG/Eu	13.1 %	43.7 %	20.0 %	63.6 %	68.8 %	49.2 %	40.9 %	32.0 %	37.5 %
HOG/χ^2	13.8 %	45.1 %	20.0 %	63.6 %	68.8 %	55.6 %	36.4 %	40.0 %	37.5 %
PHOG/Eu	16.3 %	52.1 %	60.0 %	54.5 %	68.8 %	61.9 %	54.5 %	40.0 %	37.5 %
PHOG/χ^2	15.0 %	43.7 %	60.0 %	54.5 %	68.8 %	65.1 %	18.2 %	36.0 %	43.8 %
THOG/EMD	18.8 %	49.3 %	40.0 %	63.6 %	62.5 %	55.6 %	27.3 %	40.0 %	50.0 %
THOG/$\widehat{\text{EMD}}$	8.8 %	60.6 %	40.0 %	54.5 %	50.0 %	50.8 %	27.3 %	24.0 %	50.0 %
SIFT/EMD	6.3 %	57.7 %	20.0 %	27.3 %	43.8 %	57.1 %	9.1 %	28.0 %	50.0 %
SURF/EMD	7.5 %	49.3 %	40.0 %	36.4 %	43.8 %	57.1 %	59.1 %	40.0 %	50.0 %
SIFT/$\widehat{\text{EMD}}$	11.3 %	59.2 %	60.0 %	54.5 %	25.0 %	55.6 %	31.8 %	24.0 %	43.8 %
SURF/$\widehat{\text{EMD}}$	5.6 %	47.9 %	60.0 %	45.5 %	37.5 %	33.3 %	72.7 %	56.0 %	37.5 %
Majority vote	14.4 %	47.9 %	20.0 %	63.6 %	62.5 %	55.6 %	22.7 %	24.0 %	37.5 %

F.5 Contour/Chamfer/Gradient descent

Feature/Metric	all 12	1001/1201	1011/1031	1301/1521	1311/1341	1001/1130	1511/1521	1341/1511	1110/1601
Dummy	26.3 %	59.2 %	60.0 %	81.8 %	68.8 %	66.7 %	90.9 %	80.0 %	56.3 %
Contour/Cf	21.3 %	64.8 %	60.0 %	63.6 %	50.0 %	61.9 %	13.6 %	40.0 %	50.0 %
Canny/Cf	19.4 %	70.4 %	40.0 %	72.7 %	56.3 %	58.7 %	18.2 %	52.0 %	25.0 %
HOG/Eu	20.0 %	64.8 %	40.0 %	72.7 %	50.0 %	54.0 %	31.8 %	48.0 %	37.5 %
HOG/χ^2	19.4 %	69.0 %	40.0 %	72.7 %	50.0 %	52.4 %	36.4 %	56.0 %	37.5 %
PHOG/Eu	9.4 %	54.9 %	20.0 %	72.7 %	50.0 %	55.6 %	40.9 %	32.0 %	37.5 %
PHOG/χ^2	13.8 %	49.3 %	40.0 %	81.8 %	50.0 %	55.6 %	54.5 %	40.0 %	37.5 %
THOG/EMD	15.6 %	66.2 %	40.0 %	63.6 %	50.0 %	50.8 %	31.8 %	44.0 %	37.5 %
THOG/$\widehat{\text{EMD}}$	10.6 %	54.9 %	0.0 %	54.5 %	50.0 %	61.9 %	31.8 %	52.0 %	37.5 %
SIFT/EMD	7.5 %	45.1 %	20.0 %	36.4 %	43.8 %	46.0 %	22.7 %	24.0 %	31.3 %
SURF/EMD	11.3 %	43.7 %	40.0 %	54.5 %	43.8 %	47.6 %	40.9 %	32.0 %	37.5 %
SIFT/$\widehat{\text{EMD}}$	8.8 %	69.0 %	40.0 %	45.5 %	43.8 %	50.8 %	36.4 %	44.0 %	62.5 %
SURF/$\widehat{\text{EMD}}$	6.3 %	50.7 %	0.0 %	36.4 %	56.3 %	44.4 %	68.2 %	68.0 %	43.8 %
Majority vote	22.5 %	67.6 %	40.0 %	81.8 %	43.8 %	60.3 %	27.3 %	32.0 %	37.5 %

F.6 Canny/Chamfer/None (45°)

Feature/Metric	all 12	1001/1201	1011/1031	1301/1521	1311/1341	1001/1130	1511/1521	1341/1511	1110/1601
Dummy	26.3 %	59.2 %	60.0 %	81.8 %	68.8 %	66.7 %	90.9 %	80.0 %	56.3 %
Contour/Cf	15.0 %	57.7 %	40.0 %	18.2 %	43.8 %	58.7 %	9.1 %	20.0 %	31.3 %
Canny/Cf	8.1 %	52.1 %	40.0 %	54.5 %	68.8 %	57.1 %	9.1 %	20.0 %	37.5 %
HOG/Eu	11.9 %	56.3 %	40.0 %	54.5 %	68.8 %	41.3 %	22.7 %	28.0 %	37.5 %
HOG/χ^2	11.3 %	56.3 %	40.0 %	45.5 %	68.8 %	44.4 %	27.3 %	20.0 %	37.5 %
PHOG/Eu	13.1 %	46.5 %	40.0 %	72.7 %	68.8 %	58.7 %	27.3 %	56.0 %	43.8 %
PHOG/χ^2	11.3 %	53.5 %	60.0 %	72.7 %	68.8 %	61.9 %	31.8 %	44.0 %	56.3 %
THOG/EMD	12.5 %	57.7 %	40.0 %	54.5 %	62.5 %	58.7 %	22.7 %	32.0 %	50.0 %
THOG/$\widehat{\text{EMD}}$	10.0 %	53.5 %	40.0 %	63.6 %	68.8 %	46.0 %	27.3 %	20.0 %	31.3 %
SIFT/EMD	8.8 %	43.7 %	40.0 %	54.5 %	56.3 %	55.6 %	36.4 %	44.0 %	50.0 %
SURF/EMD	8.1 %	42.3 %	40.0 %	54.5 %	81.3 %	55.6 %	50.0 %	40.0 %	43.8 %
SIFT/$\widehat{\text{EMD}}$	10.6 %	63.4 %	60.0 %	27.3 %	25.0 %	55.6 %	54.5 %	44.0 %	31.3 %
SURF/$\widehat{\text{EMD}}$	7.5 %	49.3 %	60.0 %	54.5 %	50.0 %	36.5 %	36.4 %	24.0 %	25.0 %
Majority vote	11.3 %	53.5 %	40.0 %	63.6 %	62.5 %	58.7 %	9.1 %	20.0 %	37.5 %

F.7 Canny/Chamfer/None (30°)

Feature/Metric	all 12	1001/1201	1011/1031	1301/1521	1311/1341	1001/1130	1511/1521	1341/1511	1110/1601
Dummy	26.3 %	59.2 %	60.0 %	81.8 %	68.8 %	66.7 %	90.9 %	80.0 %	56.3 %
Contour/Cf	13.8 %	50.7 %	40.0 %	45.5 %	56.3 %	58.7 %	13.6 %	28.0 %	37.5 %
Canny/Cf	5.6 %	56.3 %	40.0 %	45.5 %	68.8 %	57.1 %	9.1 %	20.0 %	50.0 %
HOG/Eu	12.5 %	47.9 %	40.0 %	72.7 %	62.5 %	55.6 %	13.6 %	36.0 %	43.8 %
HOG/χ^2	14.4 %	50.7 %	40.0 %	72.7 %	62.5 %	57.1 %	13.6 %	28.0 %	37.5 %
PHOG/Eu	11.3 %	60.6 %	20.0 %	72.7 %	62.5 %	61.9 %	31.8 %	52.0 %	37.5 %
PHOG/χ^2	13.1 %	53.5 %	20.0 %	72.7 %	62.5 %	63.5 %	13.6 %	52.0 %	43.8 %
THOG/EMD	13.8 %	53.5 %	20.0 %	63.6 %	68.8 %	58.7 %	22.7 %	32.0 %	43.8 %
THOG/$\widehat{\text{EMD}}$	11.3 %	50.7 %	40.0 %	72.7 %	56.3 %	60.3 %	22.7 %	28.0 %	25.0 %
SIFT/EMD	7.5 %	49.3 %	60.0 %	45.5 %	56.3 %	49.2 %	31.8 %	52.0 %	62.5 %
SURF/EMD	8.1 %	42.3 %	0.0 %	36.4 %	62.5 %	52.4 %	27.3 %	16.0 %	37.5 %
SIFT/$\widehat{\text{EMD}}$	7.5 %	45.1 %	20.0 %	36.4 %	43.8 %	54.0 %	36.4 %	56.0 %	50.0 %
SURF/$\widehat{\text{EMD}}$	10.0 %	43.7 %	60.0 %	54.5 %	37.5 %	38.1 %	40.9 %	36.0 %	50.0 %
Majority vote	13.8 %	53.5 %	40.0 %	63.6 %	68.8 %	57.1 %	13.6 %	28.0 %	37.5 %

F.8 Canny/Chamfer/None (22.5°)

Feature/Metric	all 12	1001/1201	1011/1031	1301/1521	1311/1341	1001/1130	1511/1521	1341/1511	1110/1601
Dummy	26.3 %	59.2 %	60.0 %	81.8 %	68.8 %	66.7 %	90.9 %	80.0 %	56.3 %
Contour/Cf	10.6 %	54.9 %	40.0 %	54.5 %	62.5 %	55.6 %	18.2 %	28.0 %	43.8 %
Canny/Cf	6.3 %	50.7 %	40.0 %	36.4 %	62.5 %	60.3 %	9.1 %	24.0 %	43.8 %
HOG/Eu	12.5 %	54.9 %	40.0 %	45.5 %	62.5 %	57.1 %	40.9 %	36.0 %	43.8 %
HOG/χ^2	15.6 %	54.9 %	40.0 %	54.5 %	56.3 %	60.3 %	31.8 %	36.0 %	43.8 %
PHOG/Eu	11.9 %	52.1 %	60.0 %	72.7 %	56.3 %	50.8 %	50.0 %	44.0 %	68.8 %
PHOG/χ^2	13.8 %	57.7 %	60.0 %	72.7 %	56.3 %	54.0 %	27.3 %	24.0 %	56.3 %
THOG/EMD	13.8 %	54.9 %	40.0 %	63.6 %	56.3 %	50.8 %	18.2 %	28.0 %	37.5 %
THOG/$\widehat{\text{EMD}}$	12.5 %	59.2 %	40.0 %	54.5 %	75.0 %	55.6 %	22.7 %	24.0 %	37.5 %
SIFT/EMD	10.0 %	53.5 %	40.0 %	27.3 %	68.8 %	57.1 %	45.5 %	60.0 %	43.8 %
SURF/EMD	3.8 %	43.7 %	40.0 %	63.6 %	75.0 %	50.8 %	36.4 %	20.0 %	50.0 %
SIFT/$\widehat{\text{EMD}}$	12.5 %	57.7 %	40.0 %	54.5 %	50.0 %	52.4 %	40.9 %	52.0 %	50.0 %
SURF/$\widehat{\text{EMD}}$	10.0 %	49.3 %	20.0 %	63.6 %	62.5 %	46.0 %	45.5 %	44.0 %	37.5 %
Majority vote	15.0 %	57.7 %	40.0 %	54.5 %	62.5 %	63.5 %	27.3 %	24.0 %	43.8 %

F.9 Canny/Chamfer/Gradient descent

Feature/Metric	all 12	1001/1201	1011/1031	1301/1521	1311/1341	1001/1130	1511/1521	1341/1511	1110/1601
Dummy	26.3 %	59.2 %	60.0 %	81.8 %	68.8 %	66.7 %	90.9 %	80.0 %	56.3 %
Contour/Cf	20.0 %	78.9 %	20.0 %	54.5 %	62.5 %	65.1 %	27.3 %	40.0 %	25.0 %
Canny/Cf	21.3 %	64.8 %	40.0 %	72.7 %	56.3 %	65.1 %	13.6 %	48.0 %	43.8 %
HOG/Eu	18.8 %	69.0 %	40.0 %	81.8 %	43.8 %	60.3 %	40.9 %	60.0 %	37.5 %
HOG/χ^2	18.8 %	71.8 %	40.0 %	81.8 %	50.0 %	68.3 %	40.9 %	60.0 %	43.8 %
PHOG/Eu	13.8 %	54.9 %	20.0 %	81.8 %	56.3 %	54.0 %	45.5 %	28.0 %	37.5 %
PHOG/χ^2	14.4 %	60.6 %	20.0 %	81.8 %	56.3 %	60.3 %	31.8 %	32.0 %	18.8 %
THOG/EMD	16.3 %	69.0 %	40.0 %	81.8 %	43.8 %	69.8 %	31.8 %	48.0 %	31.3 %
THOG/$\widehat{\text{EMD}}$	7.5 %	50.7 %	60.0 %	36.4 %	56.3 %	54.0 %	13.6 %	36.0 %	50.0 %
SIFT/EMD	7.5 %	39.4 %	20.0 %	27.3 %	56.3 %	41.3 %	18.2 %	24.0 %	18.8 %
SURF/EMD	7.5 %	46.5 %	80.0 %	27.3 %	25.0 %	49.2 %	36.4 %	36.0 %	62.5 %
SIFT/$\widehat{\text{EMD}}$	11.3 %	49.3 %	60.0 %	45.5 %	56.3 %	57.1 %	31.8 %	48.0 %	25.0 %
SURF/$\widehat{\text{EMD}}$	7.5 %	49.3 %	40.0 %	81.8 %	25.0 %	46.0 %	72.7 %	68.0 %	43.8 %
Majority vote	19.4 %	67.6 %	40.0 %	81.8 %	62.5 %	66.7 %	22.7 %	40.0 %	25.0 %

F.10 HOG/Euclidean/None (45°)

Feature/Metric	all 12	1001/1201	1011/1031	1301/1521	1311/1341	1001/1130	1511/1521	1341/1511	1110/1601
Dummy	26.3 %	59.2 %	60.0 %	81.8 %	68.8 %	66.7 %	90.9 %	80.0 %	56.3 %
Contour/Cf	12.5 %	52.1 %	40.0 %	45.5 %	75.0 %	50.8 %	13.6 %	24.0 %	43.8 %
Canny/Cf	11.3 %	56.3 %	40.0 %	45.5 %	75.0 %	58.7 %	13.6 %	16.0 %	25.0 %
HOG/Eu	8.8 %	53.5 %	40.0 %	81.8 %	68.8 %	42.9 %	36.4 %	40.0 %	37.5 %
HOG/χ^2	8.8 %	54.9 %	20.0 %	72.7 %	68.8 %	41.3 %	22.7 %	24.0 %	37.5 %
PHOG/Eu	15.0 %	56.3 %	60.0 %	63.6 %	75.0 %	65.1 %	18.2 %	28.0 %	31.3 %
PHOG/χ^2	16.9 %	54.9 %	60.0 %	63.6 %	81.3 %	65.1 %	13.6 %	24.0 %	31.3 %
THOG/EMD	13.8 %	56.3 %	20.0 %	45.5 %	68.8 %	57.1 %	13.6 %	24.0 %	43.8 %
THOG/$\widehat{\text{EMD}}$	10.0 %	43.7 %	40.0 %	45.5 %	81.3 %	41.3 %	36.4 %	24.0 %	25.0 %
SIFT/EMD	10.0 %	47.9 %	40.0 %	45.5 %	75.0 %	49.2 %	40.9 %	48.0 %	50.0 %
SURF/EMD	5.6 %	50.7 %	60.0 %	54.5 %	56.3 %	57.1 %	13.6 %	28.0 %	50.0 %
SIFT/$\widehat{\text{EMD}}$	11.3 %	45.1 %	60.0 %	36.4 %	25.0 %	57.1 %	36.4 %	56.0 %	37.5 %
SURF/$\widehat{\text{EMD}}$	7.5 %	33.8 %	60.0 %	54.5 %	50.0 %	34.9 %	27.3 %	24.0 %	43.8 %
Majority vote	11.9 %	52.1 %	60.0 %	72.7 %	75.0 %	57.1 %	18.2 %	20.0 %	43.8 %

F.11 HOG/Euclidean/None (30°)

Feature/Metric	all 12	1001/1201	1011/1031	1301/1521	1311/1341	1001/1130	1511/1521	1341/1511	1110/1601
Dummy	26.3 %	59.2 %	60.0 %	81.8 %	68.8 %	66.7 %	90.9 %	80.0 %	56.3 %
Contour/Cf	11.9 %	52.1 %	40.0 %	63.6 %	50.0 %	54.0 %	22.7 %	28.0 %	31.3 %
Canny/Cf	9.4 %	57.7 %	40.0 %	45.5 %	56.3 %	58.7 %	18.2 %	28.0 %	31.3 %
HOG/Eu	10.0 %	54.9 %	40.0 %	63.6 %	68.8 %	57.1 %	13.6 %	28.0 %	37.5 %
HOG/χ^2	12.5 %	59.2 %	40.0 %	63.6 %	68.8 %	58.7 %	13.6 %	28.0 %	31.3 %
PHOG/Eu	16.3 %	54.9 %	40.0 %	72.7 %	50.0 %	68.3 %	27.3 %	32.0 %	56.3 %
PHOG/χ^2	15.0 %	59.2 %	40.0 %	63.6 %	62.5 %	69.8 %	13.6 %	24.0 %	50.0 %
THOG/EMD	16.9 %	54.9 %	40.0 %	54.5 %	50.0 %	58.7 %	27.3 %	32.0 %	43.8 %
THOG/$\widehat{\text{EMD}}$	11.3 %	42.3 %	40.0 %	36.4 %	62.5 %	49.2 %	22.7 %	28.0 %	43.8 %
SIFT/EMD	3.1 %	50.7 %	60.0 %	45.5 %	62.5 %	54.0 %	31.8 %	32.0 %	50.0 %
SURF/EMD	11.9 %	54.9 %	20.0 %	27.3 %	25.0 %	61.9 %	36.4 %	28.0 %	56.3 %
SIFT/$\widehat{\text{EMD}}$	11.3 %	54.9 %	60.0 %	45.5 %	50.0 %	57.1 %	40.9 %	40.0 %	37.5 %
SURF/$\widehat{\text{EMD}}$	10.0 %	38.0 %	40.0 %	63.6 %	31.3 %	39.7 %	50.0 %	48.0 %	37.5 %
Majority vote	15.6 %	54.9 %	40.0 %	54.5 %	62.5 %	68.3 %	13.6 %	24.0 %	25.0 %

F.12 HOG/Euclidean/None (22.5°)

Feature/Metric	all 12	1001/1201	1011/1031	1301/1521	1311/1341	1001/1130	1511/1521	1341/1511	1110/1601
Dummy	26.3 %	59.2 %	60.0 %	81.8 %	68.8 %	66.7 %	90.9 %	80.0 %	56.3 %
Contour/Cf	13.1 %	43.7 %	40.0 %	63.6 %	62.5 %	54.0 %	22.7 %	28.0 %	31.3 %
Canny/Cf	8.1 %	52.1 %	40.0 %	54.5 %	62.5 %	58.7 %	22.7 %	24.0 %	56.3 %
HOG/Eu	11.3 %	47.9 %	40.0 %	63.6 %	68.8 %	57.1 %	18.2 %	32.0 %	37.5 %
HOG/χ^2	11.9 %	53.5 %	40.0 %	72.7 %	68.8 %	57.1 %	22.7 %	28.0 %	37.5 %
PHOG/Eu	13.8 %	47.9 %	40.0 %	63.6 %	62.5 %	63.5 %	27.3 %	32.0 %	56.3 %
PHOG/χ^2	14.4 %	50.7 %	20.0 %	72.7 %	62.5 %	60.3 %	13.6 %	36.0 %	62.5 %
THOG/EMD	14.4 %	43.7 %	20.0 %	63.6 %	68.8 %	49.2 %	22.7 %	24.0 %	37.5 %
THOG/$\widehat{\text{EMD}}$	11.3 %	38.0 %	40.0 %	63.6 %	62.5 %	39.7 %	36.4 %	32.0 %	43.8 %
SIFT/EMD	6.3 %	47.9 %	40.0 %	36.4 %	62.5 %	54.0 %	22.7 %	48.0 %	56.3 %
SURF/EMD	9.4 %	53.5 %	40.0 %	36.4 %	56.3 %	61.9 %	27.3 %	20.0 %	43.8 %
SIFT/$\widehat{\text{EMD}}$	8.8 %	53.5 %	20.0 %	45.5 %	56.3 %	47.6 %	27.3 %	36.0 %	37.5 %
SURF/$\widehat{\text{EMD}}$	7.5 %	54.9 %	40.0 %	45.5 %	31.3 %	44.4 %	36.4 %	24.0 %	43.8 %
Majority vote	15.6 %	46.5 %	40.0 %	63.6 %	62.5 %	58.7 %	18.2 %	20.0 %	37.5 %

F.13 HOG/Euclidean/Gradient descent

Feature/Metric	all 12	1001/1201	1011/1031	1301/1521	1311/1341	1001/1130	1511/1521	1341/1511	1110/1601
Dummy	26.3 %	59.2 %	60.0 %	81.8 %	68.8 %	66.7 %	90.9 %	80.0 %	56.3 %
Contour/Cf	23.1 %	67.6 %	60.0 %	72.7 %	75.0 %	68.3 %	22.7 %	36.0 %	31.3 %
Canny/Cf	20.6 %	71.8 %	60.0 %	63.6 %	50.0 %	71.4 %	22.7 %	40.0 %	50.0 %
HOG/Eu	26.3 %	77.5 %	60.0 %	90.9 %	68.8 %	82.5 %	18.2 %	60.0 %	43.8 %
HOG/χ^2	27.5 %	77.5 %	60.0 %	90.9 %	68.8 %	76.2 %	18.2 %	56.0 %	37.5 %
PHOG/Eu	10.0 %	63.4 %	40.0 %	90.9 %	68.8 %	63.5 %	40.9 %	48.0 %	37.5 %
PHOG/χ^2	15.0 %	64.8 %	40.0 %	81.8 %	68.8 %	68.3 %	27.3 %	40.0 %	31.3 %
THOG/EMD	21.9 %	64.8 %	60.0 %	72.7 %	75.0 %	65.1 %	22.7 %	40.0 %	50.0 %
THOG/$\widehat{\text{EMD}}$	19.4 %	62.0 %	80.0 %	54.5 %	75.0 %	65.1 %	27.3 %	36.0 %	50.0 %
SIFT/EMD	8.1 %	53.5 %	20.0 %	36.4 %	56.3 %	50.8 %	18.2 %	32.0 %	31.3 %
SURF/EMD	8.1 %	52.1 %	80.0 %	36.4 %	50.0 %	50.8 %	27.3 %	28.0 %	43.8 %
SIFT/$\widehat{\text{EMD}}$	11.3 %	56.3 %	60.0 %	54.5 %	56.3 %	50.8 %	36.4 %	28.0 %	31.3 %
SURF/$\widehat{\text{EMD}}$	7.5 %	36.6 %	60.0 %	63.6 %	31.3 %	39.7 %	50.0 %	56.0 %	18.8 %
Majority vote	22.5 %	71.8 %	60.0 %	81.8 %	68.8 %	73.0 %	13.6 %	40.0 %	31.3 %

F.14 HOG/Euclidean/L-BFGS

Feature/Metric	all 12	1001/1201	1011/1031	1301/1521	1311/1341	1001/1130	1511/1521	1341/1511	1110/1601
Dummy	26.3 %	59.2 %	60.0 %	81.8 %	68.8 %	66.7 %	90.9 %	80.0 %	56.3 %
Contour/Cf	11.9 %	43.7 %	40.0 %	36.4 %	25.0 %	58.7 %	13.6 %	12.0 %	0.0 %
Canny/Cf	10.6 %	42.3 %	40.0 %	36.4 %	18.8 %	58.7 %	13.6 %	12.0 %	0.0 %
HOG/Eu	8.8 %	40.8 %	40.0 %	45.5 %	25.0 %	60.3 %	4.5 %	4.0 %	0.0 %
HOG/χ^2	10.6 %	49.3 %	40.0 %	45.5 %	25.0 %	60.3 %	4.5 %	8.0 %	0.0 %
PHOG/Eu	13.1 %	38.0 %	40.0 %	36.4 %	25.0 %	61.9 %	13.6 %	12.0 %	0.0 %
PHOG/χ^2	13.8 %	39.4 %	40.0 %	36.4 %	25.0 %	63.5 %	13.6 %	12.0 %	0.0 %
THOG/EMD	11.3 %	40.8 %	40.0 %	36.4 %	25.0 %	55.6 %	13.6 %	8.0 %	0.0 %
THOG/$\widehat{\text{EMD}}$	11.9 %	40.8 %	40.0 %	36.4 %	18.8 %	58.7 %	13.6 %	12.0 %	0.0 %
SIFT/EMD	11.3 %	47.9 %	40.0 %	18.2 %	12.5 %	55.6 %	9.1 %	12.0 %	0.0 %
SURF/EMD	13.1 %	45.1 %	40.0 %	36.4 %	6.3 %	58.7 %	18.2 %	16.0 %	0.0 %
SIFT/$\widehat{\text{EMD}}$	13.1 %	57.7 %	40.0 %	18.2 %	6.3 %	55.6 %	13.6 %	12.0 %	0.0 %
SURF/$\widehat{\text{EMD}}$	9.4 %	38.0 %	40.0 %	36.4 %	18.8 %	65.1 %	13.6 %	12.0 %	0.0 %
Majority vote	11.9 %	42.3 %	40.0 %	36.4 %	25.0 %	60.3 %	13.6 %	12.0 %	0.0 %

F.15 Trimmed HOG (size 50)/$\widehat{\text{EMD}}$/None ($45°$)

Feature/Metric	all 12	1001/1201	1011/1031	1301/1521	1311/1341	1001/1130	1511/1521	1341/1511	1110/1601
Dummy	27.1 %	59.2 %	50.0 %	81.8 %	68.8 %	67.7 %	90.0 %	78.3 %	60.0 %
Contour/Cf	11.6 %	50.7 %	50.0 %	27.3 %	75.0 %	43.5 %	25.0 %	30.4 %	33.3 %
Canny/Cf	12.3 %	49.3 %	50.0 %	63.6 %	75.0 %	48.4 %	25.0 %	26.1 %	53.3 %
HOG/Eu	11.0 %	47.9 %	75.0 %	63.6 %	81.3 %	45.2 %	50.0 %	47.8 %	40.0 %
HOG/χ^2	11.0 %	46.5 %	75.0 %	72.7 %	81.3 %	43.5 %	35.0 %	43.5 %	40.0 %
PHOG/Eu	11.0 %	43.7 %	100.0 %	45.5 %	68.8 %	48.4 %	45.0 %	47.8 %	73.3 %
PHOG/χ^2	10.3 %	47.9 %	100.0 %	45.5 %	81.3 %	51.6 %	45.0 %	47.8 %	60.0 %
THOG/EMD	9.0 %	47.9 %	75.0 %	36.4 %	75.0 %	48.4 %	25.0 %	30.4 %	26.7 %
THOG/$\widehat{\text{EMD}}$	11.6 %	46.5 %	75.0 %	63.6 %	68.8 %	45.2 %	35.0 %	30.4 %	46.7 %
SIFT/EMD	7.1 %	50.7 %	25.0 %	27.3 %	68.8 %	51.6 %	30.0 %	52.2 %	46.7 %
SURF/EMD	1.9 %	43.7 %	75.0 %	45.5 %	25.0 %	50.0 %	45.0 %	34.8 %	33.3 %
SIFT/$\widehat{\text{EMD}}$	8.4 %	60.6 %	50.0 %	36.4 %	31.3 %	53.2 %	40.0 %	56.5 %	40.0 %
SURF/$\widehat{\text{EMD}}$	8.4 %	39.4 %	75.0 %	54.5 %	43.8 %	35.5 %	50.0 %	30.4 %	20.0 %
Majority vote	11.6 %	43.7 %	75.0 %	54.5 %	75.0 %	43.5 %	30.0 %	34.8 %	40.0 %

F.16 Trimmed HOG (size 50)/$\widehat{\text{EMD}}$/None (30°)

Feature/Metric	all 12	1001/1201	1011/1031	1301/1521	1311/1341	1001/1130	1511/1521	1341/1511	1110/1601
Dummy	27.1 %	59.2 %	50.0 %	81.8 %	68.8 %	67.7 %	90.0 %	78.3 %	60.0 %
Contour/Cf	12.3 %	46.5 %	50.0 %	36.4 %	62.5 %	56.5 %	15.0 %	30.4 %	46.7 %
Canny/Cf	9.0 %	46.5 %	50.0 %	45.5 %	56.3 %	59.7 %	20.0 %	39.1 %	33.3 %
HOG/Eu	12.9 %	57.7 %	50.0 %	36.4 %	56.3 %	54.8 %	20.0 %	43.5 %	40.0 %
HOG/χ^2	13.5 %	54.9 %	50.0 %	36.4 %	56.3 %	51.6 %	20.0 %	43.5 %	40.0 %
PHOG/Eu	8.4 %	50.7 %	25.0 %	36.4 %	43.8 %	46.8 %	40.0 %	43.5 %	46.7 %
PHOG/χ^2	10.3 %	50.7 %	25.0 %	36.4 %	56.3 %	58.1 %	30.0 %	39.1 %	40.0 %
THOG/EMD	11.0 %	47.9 %	50.0 %	54.5 %	75.0 %	50.0 %	25.0 %	26.1 %	46.7 %
THOG/$\widehat{\text{EMD}}$	10.3 %	53.5 %	75.0 %	54.5 %	56.3 %	43.5 %	25.0 %	43.5 %	33.3 %
SIFT/EMD	9.7 %	46.5 %	50.0 %	27.3 %	62.5 %	56.5 %	40.0 %	52.2 %	46.7 %
SURF/EMD	11.0 %	46.5 %	50.0 %	36.4 %	37.5 %	45.2 %	40.0 %	30.4 %	46.7 %
SIFT/$\widehat{\text{EMD}}$	12.9 %	46.5 %	25.0 %	45.5 %	37.5 %	59.7 %	55.0 %	52.2 %	60.0 %
SURF/$\widehat{\text{EMD}}$	7.7 %	47.9 %	25.0 %	45.5 %	37.5 %	48.4 %	75.0 %	56.5 %	46.7 %
Majority vote	12.3 %	50.7 %	50.0 %	27.3 %	56.3 %	51.6 %	25.0 %	30.4 %	40.0 %

F.17 Trimmed HOG (size 50)/$\widehat{\text{EMD}}$/None (22.5°)

Feature/Metric	all 12	1001/1201	1011/1031	1301/1521	1311/1341	1001/1130	1511/1521	1341/1511	1110/1601
Dummy	27.1 %	59.2 %	50.0 %	81.8 %	68.8 %	67.7 %	90.0 %	78.3 %	60.0 %
Contour/Cf	11.0 %	56.3 %	50.0 %	81.8 %	68.8 %	66.1 %	20.0 %	30.4 %	46.7 %
Canny/Cf	9.0 %	54.9 %	50.0 %	90.9 %	56.3 %	64.5 %	15.0 %	30.4 %	46.7 %
HOG/Eu	10.3 %	49.3 %	50.0 %	72.7 %	68.8 %	56.5 %	20.0 %	39.1 %	53.3 %
HOG/χ^2	10.3 %	52.1 %	50.0 %	72.7 %	68.8 %	58.1 %	20.0 %	30.4 %	53.3 %
PHOG/Eu	12.3 %	63.4 %	50.0 %	81.8 %	75.0 %	56.5 %	15.0 %	39.1 %	46.7 %
PHOG/χ^2	16.8 %	56.3 %	25.0 %	72.7 %	75.0 %	64.5 %	15.0 %	47.8 %	53.3 %
THOG/EMD	17.4 %	64.8 %	50.0 %	72.7 %	75.0 %	61.3 %	40.0 %	43.5 %	46.7 %
THOG/$\widehat{\text{EMD}}$	9.7 %	50.7 %	50.0 %	90.9 %	81.3 %	50.0 %	30.0 %	26.1 %	53.3 %
SIFT/EMD	10.3 %	59.2 %	25.0 %	36.4 %	50.0 %	62.9 %	30.0 %	34.8 %	40.0 %
SURF/EMD	3.9 %	47.9 %	50.0 %	27.3 %	31.3 %	51.6 %	55.0 %	39.1 %	53.3 %
SIFT/$\widehat{\text{EMD}}$	8.4 %	47.9 %	25.0 %	36.4 %	43.8 %	45.2 %	25.0 %	39.1 %	26.7 %
SURF/$\widehat{\text{EMD}}$	8.4 %	49.3 %	50.0 %	63.6 %	62.5 %	43.5 %	60.0 %	47.8 %	46.7 %
Majority vote	11.6 %	54.9 %	25.0 %	81.8 %	81.3 %	59.7 %	25.0 %	30.4 %	53.3 %

F.18 Trimmed HOG (size 50)/$\widehat{\text{EMD}}$/Gradient descent

Feature/Metric	all 12	1001/1201	1011/1031	1301/1521	1311/1341	1001/1130	1511/1521	1341/1511	1110/1601
Dummy	27.1 %	59.2 %	50.0 %	81.8 %	68.8 %	67.7 %	90.0 %	78.3 %	60.0 %
Contour/Cf	15.5 %	69.0 %	75.0 %	36.4 %	62.5 %	58.1 %	40.0 %	47.8 %	46.7 %
Canny/Cf	13.5 %	69.0 %	75.0 %	36.4 %	56.3 %	61.3 %	35.0 %	39.1 %	46.7 %
HOG/Eu	16.1 %	66.2 %	100.0 %	36.4 %	68.8 %	56.5 %	25.0 %	47.8 %	40.0 %
HOG/χ^2	15.5 %	60.6 %	50.0 %	45.5 %	68.8 %	56.5 %	25.0 %	47.8 %	33.3 %
PHOG/Eu	10.3 %	67.6 %	25.0 %	72.7 %	62.5 %	53.2 %	55.0 %	21.7 %	46.7 %
PHOG/χ^2	14.8 %	69.0 %	25.0 %	54.5 %	68.8 %	53.2 %	35.0 %	30.4 %	40.0 %
THOG/EMD	12.3 %	59.2 %	50.0 %	36.4 %	62.5 %	62.9 %	20.0 %	56.5 %	33.3 %
THOG/$\widehat{\text{EMD}}$	9.7 %	47.9 %	25.0 %	27.3 %	43.8 %	50.0 %	40.0 %	47.8 %	60.0 %
SIFT/EMD	5.2 %	49.3 %	0.0 %	45.5 %	31.3 %	50.0 %	15.0 %	21.7 %	60.0 %
SURF/EMD	8.4 %	53.5 %	50.0 %	54.5 %	31.3 %	67.7 %	35.0 %	30.4 %	66.7 %
SIFT/$\widehat{\text{EMD}}$	7.7 %	54.9 %	50.0 %	54.5 %	25.0 %	58.1 %	25.0 %	56.5 %	46.7 %
SURF/$\widehat{\text{EMD}}$	9.0 %	45.1 %	25.0 %	72.7 %	56.3 %	51.6 %	45.0 %	60.9 %	53.3 %
Majority vote	12.9 %	62.0 %	50.0 %	36.4 %	50.0 %	54.8 %	25.0 %	34.8 %	40.0 %