A COMPOSITIONAL APPROACH TO HAND GESTURE RECOGNITION

Teză destinată obţinerii
titlului ştiinţific de doctor inginer
la
Universitatea "Politehnica" din Timişoara
în domeniul INGINERIE ELECTRONICĂ ŞI TELECOMUNICAŢII
de către

Ing. Georgiana Sârbu-Doagă

Conducător ştiinţific: prof.univ.dr.ing.Marius Oteşteanu
Referenţi ştiinţifici: prof.univ.dr.ing.Victor Neagoe
prof.univ.dr.ing.Franz Quint
prof.univ.dr.ing.Vasile Gui

Ziua susţinerii tezei: 23.09.2009

Seriile Teze de doctorat ale UPT sunt:

1. Automatică
2. Chimie
3. Energetică
4. Ingineria Chimică
5. Inginerie Civilă
6. Inginerie Electrică
7. Inginerie Electronică şi Telecomunicaţii
8. Inginerie Industrială
9. Inginerie Mecanică
10. Ştiinţa Calculatoarelor
11. Ştiinţa şi Ingineria Materialelor

Universitatea „Politehnica” din Timişoara a iniţiat seriile de mai sus în scopul diseminării expertizei, cunoştinţelor şi rezultatelor cercetărilor întreprinse în cadrul şcolii doctorale a universităţii. Seriile conţin, potrivit H.B.Ex.S Nr. 14 / 14.07.2006, tezele de doctorat susţinute în universitate începând cu 1 octombrie 2006.

România, 300159 Timişoara, Bd. Republicii 9,
tel. 0256 403823, fax. 0256 403221
e-mail: editura@edipol.upt.ro

Acknowledgments

I have worked on this thesis during my research in the Communication Department of "Politehnica" University, Timişoara, and it is the result of a sustained effort over the last 4 years.

The goal of this work was to prove the power of compositional techniques in hand gesture recognition. The main advantage of the compositional techniques is their generality; these techniques are more independent of application. Compositional techniques are well suited to incorporate principles from the Gestalt theory of visual perception, therefore they have an important and mostly unexplored potential for further development.

My sincere thanks go to all people that have contributed in various ways to the successful completion of this thesis. First I would like to thank to my supervisor Professor Marius Otesteanu and to Professor Vasile Gui for their support and guidance.

Special thanks go to my family for their love, patience and encouragement.

Timişoara, September 2009 Georgiana Sârbu-Doagă

Destinatarii dedicaţiei.

Georgiana, Sârbu-Doagă

A Compositional Approach to Hand Gesture Recognition.

Teze de doctorat ale UPT, Seria 7, Nr. 18, Editura Politehnica, 2009, 134 pagini, 78 figuri, 17 tabele.

ISSN: 1842-7014

ISBN: 978-973-625-956-2

Key words: hand gesture recognition, compositional, sparse representations, Harris corner detector, clustering.

Abstract,

The direct use of the hand as an input device is an attractive method for providing natural human–computer interaction. Computer Vision community tried to solve the hand gesture recognition problem by following the paradigm: hand localization, hand segmentation, feature extraction and next step hand classification, without any abstract representation in between the last two steps. In this work is proposed to use compositional techniques in order to recognize the hand gestures. Compositional representations split complex objects into simpler parts, which are easier to recognize and using the relationships between them, the complex object is recognized. The main advantage of the compositional techniques is their generality; these techniques are more independent of application. Using these techniques we address also to the semantic gap that exists between the low level features and high level representations.

This work is an attempt to extend the types of problems solved based on the new, compositional approach. The hand posture representation is based on *compositions* of parts: descriptors are grouped according to the perceptual laws of grouping obtain a set of possible candidate compositions. These groups are a sparse representation of the hand posture based on overlapping subregions.

The power of compositional techniques for hand gesture recognition is proved by the results that have been obtained.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 MOTIVATION

People perform various gestures in their daily lives. It is in our nature to use gestures in order to improve the communication between us and the people that surround us. Try to imagine speaking with a person who makes no gesture, someone who is not moving his or her hands or has no has facial expression. It is very difficult to understand if your message is clear for him or her, if he or she agrees with your saying, in other words it is very hard to guess what type of reaction your message produces. Between all kind of gestures that we perform, hand gestures play an important role. Hand gestures can help us say more in less time, if we use them correctly. By using just a hand gesture it is possible to say: good luck, I am watching you, I love you, patience, hello, good bye, quiet, ask for money, time out and so one. In these days, computers have become an important part in our lives, so why not use hand gesture in order to communicate with them.

1.1 The problem

The direct use of the hand as an input device is an attractive method for providing natural human–computer interaction (HCI). Since now, the only technology that satisfies the advanced requirements of hand-based input for HCI is glove-based sensing. Several drawbacks make this technology not so popular: first of all interaction with the computer-controlled environment loses naturalness and easiness and it also requires calibration and setup procedures. Computer vision has the potential to provide more natural and non-contact solutions, but has no lake of challenges including accuracy, processing speed, and general have to be overcome for the widespread use of this technology. Vision based models can be classified in two groups: 3D model based and appearance based models. The 3D hand models are articulated deformable objects with many degrees of freedom; a very large image database is required to cover all the characteristic shapes under different views. Another common problem with model based approaches is the problem of feature extraction and lack of capability to deal with singularities that arise from ambiguous views. By using appearance based methods, computer vision community tried to solve the hand gesture recognition problem by following the paradigm: hand localization, hand segmentation, feature extraction and next step hand classification, without any abstract representation in between the last two steps. View-based methods have been shown to be effective at discriminating between a certain number of hand poses, which is satisfactory for a number of applications in gesture recognition. One of the main problems in view-based methods remains the segmentation stage.

1.2 The Solution

Compared to traditional appearance based approaches, this work proposes to use compositional techniques in order to recognize the hand gestures.

Why compositional techniques may one wonder? Because these techniques are able to emulate better the way people think. Even if my field of interest was wider [1], [2], [3], [4], [5], I narrowed it down to compositional techniques [6], [7], [8], [9]. Compositionality refers to the prominent ability of human cognition to represent entities as hierarchies of meaningful and generic parts.

Compositional representations decompose complex objects into simpler parts, which are easier to recognize and using the relationships between them, the complex object is recognized. These techniques have been studied in many diverse fields such as linguistics, logic, and neuroscience, but compositionality is especially evident in the syntax and semantics of language where a limited number of letters can form a huge variety of words and sentences. In computer vision these techniques are used in the context of a general problem: categorization. In the literature the terms class and category are often used interchangeably. Image categorization does refer to the task of deciding what category a whole image belongs to. Therefore, the image is labeled according to its most prominent object. There are also other categorization settings conceivable such as action recognition where an image or video sequence has to be labeled as featuring one of several possible actions.

The main advantage of the compositional techniques is their generality; these techniques are more independent of application. Using these techniques we address also to the semantic gap that exists between the low level features and high level representations. During the years researchers have tried to fill this gap, by using the compositional techniques there is actually a bridge build over the semantic gap. Recognition does not imply a huge step from feature to classification any more, now we can step between by building compositions. In order to build compositions the Gestalt laws of visual perception are taken into account. These laws are a set of visual rules that guide the construction process of groupings and yield compositions, establishing causal relationships between grouping constituents, and tend to emulate better the way our brain-view processor works.

The key idea of this thesis is to use these techniques for the more specific problem: the hand posture recognition. By hand posture we refer to a static hand pose without involvement of movements. This work is an attempt to extend the types of problems solved based on the new, compositional approach. While using the general framework of some reference compositional techniques [10], [11] this work designed the processing modules by considering the specifics of the hand gesture recognition problem, where needed. A hand posture representation is based on *compositions* of parts: descriptors are grouped according to the perceptual laws of grouping [12]obtain a set of possible candidate compositions. These groups are a sparse representation of the hand posture based on overlapping subregions.

The detected part descriptors are represented as probability distributions over a codebook which is obtained in the learning phase. A composition is a mixture of the part distributions. From all candidate compositions, relevant compositions must be selected. There are two types of relevant compositions: those compositions that occur frequently in all categories and also those which are specific for a category. The category posterior of compositions is learned in the training phase, and it is a measure of relevance. The entropy of the category posterior helps us to discriminate between categories. A cost function is obtained by combining the priors of the prototypes and the entropy. The process of recognition is based on bag of composition method, where a discriminative function is defined.

1.3 Contributions

The main contributions of this thesis are:

- the compositional approach used to hand posture recognition
- the careful selection of the basic features (contours, interest points, patches, colour histograms, orientation histograms), these basic features generate compositions.
- the optimizations of several parameters from the framework
- the discriminate function

1.4 Thesis Outline

This thesis is organized as follows:

- **Chapter 2** gives an overview of various crucial components of visual object recognition systems**.** In this chapter is also presented a short overview of vision based hand gesture recognition. This chapter reviews the basic concepts needed for the presentation of the composition system in later parts of the thesis.
- **Chapter 3** presents the theoretical fundamentals needed to develop the hand gesture recognition system: Canny edge detector, Harris corner detector, Principle of Compositionality, Gestalt laws, Learning Paradigms and Robust Estimation.
- **Chapter 4** presents the proposed compositional approach to hand gesture recognition. Here are presented in detail all the implementation steps of the systems starting with feature detection, object representation, training, and finally recognition. In this chapter the differences between traditional pattern recognition and compositional approaches with regard to each of these stages are pointed out. The motivations behind the modelling decision are presented.
- **Chapter 5** presents the tests results using the proposed method and prove the power of compositional techniques used in hand gesture recognition.
- **Chapter 6** presents the contributions and conclusions of this work.

2 RELATED WORK

2.1 Model-Based Vision and View-Based Approaches: An Overview

In the object recognition community, object representations have spanned a continuum ranging from generic models to appearance-based models. The evolution of object recognition according to [13] can be seen in figure 1. The rationale of model-based vision is to center the representation on the physical objects so that it becomes invariant with respect to view changes. On the other side of model-based methods, view-based approaches are founded on models which are directly centered on the observed view of an object in an image.

At the beginning of vision research image acquisition was heavy and costly, so in the 1970's, a main stream of research focused on the geometry of objects thereby leaving aside appearance information, so it can be talked about geometry without appearance. The 3-D geometry of an object used to be described by decomposing it into simple primitives. In this years vision researchers aimed for prototypical vision systems, using generalized cylinders [14], [15], [16], [17] and later super-quadrics [18], [19], [20], [21] and geons [22], [23], [24], [25].

Some examples of classical vision systems are Brooks' ACRONYM system [26] and Binford's SUCCESSOR [27].The 1980's are characterized by 3-D models that captured the exact shape of the object, this models, often in the form of CAD models, were effectively 3-D templates [28] ,[29],[30].

The model-based approaches differ from each other by the intrinsic representation they use for 3-D objects. Lowe's approach [31] is build on a direct 3-D representation while in [32] Ullman and Basiri proposed an approach to recognition in which a 3-D object is represented by the linear combination of 2-D images of the object, more precisely this approach incorporate a mixture of 2-D models that are matched to the image using lines and points.

Research of human vision has seen a related debate. Some argue that the brain uses a 3-D representation. In [33] Biederman proposes geons, a set of 3-D primitives like cones, cylinders, or cuboids, the idea is that any object is decomposed into these generic constituents; in contrast to this idea, a view-based representation has been proposed in [34].

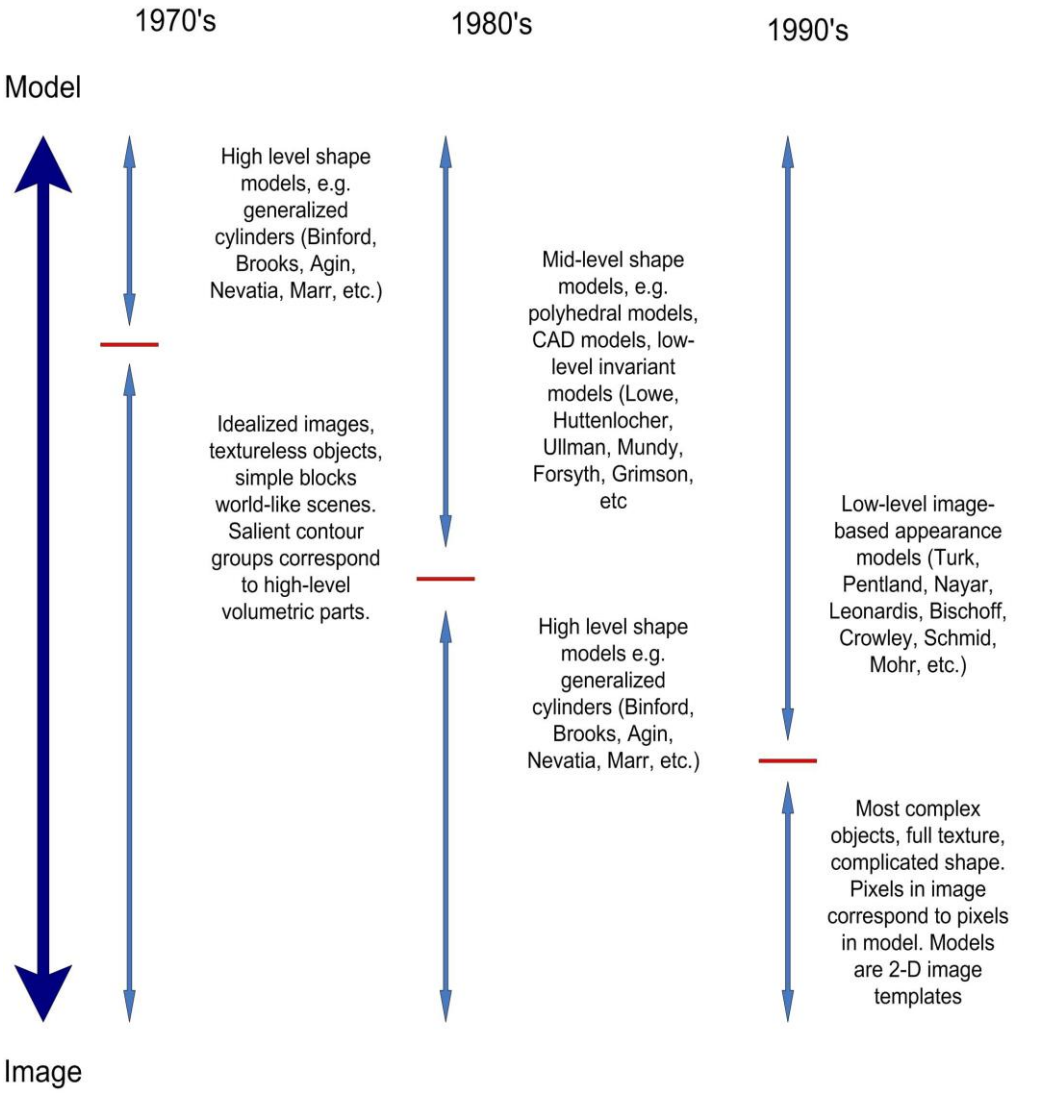


Figure 1 Evolution of object recognition

In the 1990's, appearance models have replaced CAD models, and for the first time, recognition systems were constructed that could recognize arbitrarily complex objects, [35], [36], [37].

2.1.1 Appearance without Geometry

The view-based approaches are founded on models which are directly centered on the observed view of an object in an image. In [38] a template based approach is pursued, images are sub sampled before summarizing the pixels in a vector and classification is then performed using a linear support vector machine. In [39] to obtain invariance of the object translation, is used a view-based approach with multiple detectors that are each specialized to a specific orientation of the object. In [40] to obtain invariance with respect to object translation sliding windows are used. The image is divided into regions and each of these segments is matched against the template. Based on the matching score is then decided which region contains the object of interest. Similarly, invariance to scaling, or rotation can be incorporated by additionally searching over scale and orientation. Moreover, the unsupervised dimensionality reduction methods are also popular. In order to obtain compact image description that is invariant to local image alteration (for example environmental factors like shadows or clutters), the comparison is typically conducted in the feature space, which results from projecting the image space using linear or non-linear methods. *Principle component analysis* is a widely used technique [41] to reduce the dimensionality of the data set consisting of a large number of interrelated variables while retaining the variation presented in the data set. The PCA conducted to eigenspace space representations and their particular case eigenfaces. In [35] the face images are decompose into a small set of characteristics feature images called eigenfaces. The image is recognizing by projecting it into the subspace span by the eiganfaces. In this approach each class has its own eigenspace, represented by those best eigenfaces that have the larger eigenvalues and account for the most variance within the set of face images. In [36] Murase and his team obtain a single parametric eigenspace (which encode both identity and viewing conditions) for the image set by computing the most prominent eigenvectors of the image set, then all the training sample are projected onto the eigenspace , and recognition is no more than nearest-neighbor search [42] in the eigenspace for the training closest sample. Before applying a PCA approach global image transformations such as translation, scaling, or illumination changes have to be removed in a preprocessing stage. This eigenspace approaches are holistic, the view-based models based on holistic representation are sensitive to the variation in the spatial structure of the object. The deformable template matching [43], [44] compensates for variations in the spatial structure by applying a global transformation when matching templates. The deformable models proposed in literature can be classified according to [44] in free-form and parametric. The free-form deformable models have no global structure of the template, the template constrains refers to local continuity and smoothness. The parametric form is used when it is some prior knowledge of the geometrical shape. This information is encoded in small numbers of parameters. The parameter deformable template is represented as a collection of parameterized curves or by the image of a prototype template under a parametric mapping. The model used in [44] is parametric one, the prototype template is represented as a bitmap which describe de characteristics of an object shape. The total deviation of the deformed template from the probe image should be minimized. This distance consists of the local distances between key points on the deformed template and feature points in the image. As an additional constraint, the total deformation is kept bounded so that simple transformations are preferred while minimizing local deviations between template and probe. To follow this in [45] was developed a new non-rigid point matching

algorithm which is well suited for non-rigid registration. The algorithm utilizes the softassign, deterministic annealing, the thin-plate spline for the spatial mapping and outlier rejection to solve for both the correspondence and mapping parameters. This algorithm later has served as the basis for recognition systems based on shape matching [46], [47].

A smooth move towards part based models is done in researches like [48] where special templates extract the salient image parts: eyes ,noses and mouth.

The unsupervised dimensionality reduction was not the only problem trying to be solved; discriminative techniques have also been applied in the literature. The most famous example of dimensionality reduction is PCA which on the other hand does not take into account any difference in class. PCA is an unsupervised technique and as such does not include label information of the data. The question that arises is how to utilize the label information in finding informative projections? To that purpose Fisher- linear discriminant analysis (LDA) [49], [42] considers maximizing the following objective:

$$J(w) = \frac{w^T S_B w}{w^T S_W w} \qquad (2.1)$$

where S_B is the between classes scatter matrix and S_W is the within classes scatter matrix. LDA explicitly attempts to model the difference between the classes of data. An example of LDA used in face recognition is [50].

2.1.1.1 Global Image Histograms in Content Based Image Retrieval

In content based image retrieval histograms over complete images have been popular.These methods integrate features such as color or texture over whole images. Consequently these representations are invariant with respect to local changes in an image and also to global transformations such as translation at the cost of limited specificity, sensitivity to background clutter, and limited invariance to occlusion. To enhance specificity, a smooth transition towards incorporating feature localization information (and thus global object geometry) has taken place. In this category are approaches of Swain and Ballard [51]and that of Schiele and Crowley [52]. The latter method establishes joint histograms over local appearances, which are measured by local shape descriptors. The histograms do also incorporate spatial information of the local descriptors. Vogel and Schiele [53] presented an image representation that renders it possible to access natural scenes by local semantic description. Some spatial information is incorporated by using a rigid grid of local regions. In order to recognize and detect individual objects it is necessary a representation that exhibits better localization.

2.1.1.2 Bag of features method

In the last years bag-of-words models from text retrieval have become very popular. Bag-of-words representation has proven its usefulness in text classification. Using this model, a text (such as a sentence or a document) is represented as an

order-less collection of words, disregarding grammar and even word order. Computer vision researchers have used a similar idea for image (a particular object) representation: an image can be treated as a document, and features extracted from the image are considered as the "words"; in this context they are commonly referred to as bag of features-approaches. Because the "word" in images is not the shelf thing like the word in text document, to achieve it usually are necessary the following three steps: feature detection, feature description and codebook generation. Given an image, feature detection is used to extract several local patches/regions, which are considered as candidates for basic elements, "words". One of the most simple yet effective method for feature detection is the regular grid [11], [54], the image is evenly segmented by some horizontal and vertical lines and some local patches are obtained. Using this method very promising results for natural scene categorization were obtained [11]. The interest point detectors are used to find the salient regions from an image. The salient patches such as: edges, corners and blobs are detected; these patches are considered to be more important than other patches and are more useful for object classification. The most well known detectors are: Harris corner detector [55], Lowe's Difference of Gaussians (DoG) [56] and Kadir Brady detector [57].

After feature detection, each image is abstracted by several local patches. For the detected regions remains the question how to represent the patches as numerical vectors? In order to describe the regions local descriptors are used. A good descriptor should have the ability to handle intensity, rotation, scale and affine variations. The simplest descriptor is a vector of image pixels, in praxis some extra processing is needed to reduce the dimensionality and insure invariance to at least limited image transformations. One of the most famous descriptors is Scale-invariant feature transform (SIFT) [56]. SIFT converts each patch to 128-dimensional vector. After this step, each image is a collection of vectors of the same dimension (128 for SIFT). The SIFT features are robust to changes in illumination, noise, and minor changes in viewpoint; they are highly distinctive, relatively easy to extract, allow for correct object identification with low probability of mismatch and are easy to match against a (large) database of local features. Other popular descriptors are: shape context descriptors [58], geometric blur descriptor [59], gradient location-orientation histogram (GLOH), Gabor Filters [60], steerable filters [61], complex filters [62]. The use of interest points detectors conduct to good results, this can be confirmed by the works of: [63], [64]. Moreover other methods like: random sampling [65] and segmentation [66] are used to extract the features.

These vectors which are describing the patches are used to form a codebook. Clustering is a common method for learning a visual vocabulary or codebook. In the training phase all vectors from all training images all clustered in order to form the codebook, the number of clusters is the codebook size. An image is represented as a distribution over the codebook; actually the representation is a histogram that lists the occurrence frequencies of each prototype in the image. A classifier is train to map the vectors to a class label. The main disadvantage of these methods is that they ignores the spatial relationships among the patches and capture only their co-occurrence. Their advantage is the compact representation which can be learned from a small number of samples. Over the last years several methods to incorporate the spatial information have been proposed. For discriminative models, spatial pyramid match [67] performs pyramid matching by partitioning the image into increasingly fine sub-regions and compute histograms of local features inside each sub-region. All the individual descriptors are concatenated in order to represent the image. This is close to the rigid template matching on cell

level, so this approach assumes that the spatial structure of the object is fixed whit respect to the image. For generative models, relative positions of codewords [68] are also taken into account.

Bag-of-feature approaches made popular techniques from text retrieval like latent semantic analysis (LSA). The key idea is to map high-dimensional count vectors, such as the ones arising in vector space representations of text documents [12], to a lower dimensional representation in a so-called latent semantic space. As the name suggests, the goal of LSA is to find a data mapping which provides information well beyond the lexical level and reveals semantically relations between the entities of interest. LSA can use a term-document matrix which describes the occurrences of terms in documents; it is a sparse matrix whose rows correspond to terms and whose columns correspond to documents. A typical example of the weighting of the elements of the matrix is tf-idf (term frequency–inverse document frequency): the element of the matrix is proportional to the number of times the terms appear in each document, where rare terms are up-weighted to reflect their relative importance. In image recognition each image is represented by an occurrence histogram over a fix codebook, so the set of all training samples is characterized by a large co-occurrence matrix which is then decompose using singular value decomposition. The eigenvectors correspond to different latent topics and the eigenvalues give their relative weighting. Hofmann [69] , [70] used the LSA into a probabilistic framework, this conducted to probabilistic latent semantic analysis (pLSA).

The starting point for Probabilistic Latent Semantic Analysis is a statistical model which has been called aspect model. The aspect model is a latent variable model for co-occurrence data which associates an unobserved class variable $z \in Z = \{z_1, .., z_k\}$ with each observation, for example with each occurrence of a word $w \in W = \{w_1, ..., w_M\}$ in a document $d \in D = \{d_1, ..., d_N\}$. The whole corpus of samples is described by $W \times D$ co-occurrence matrix of joint probabilities $P(w, d)$. The pLSA introduces a new hidden layer of latent topics $z \in Z$ that d-separate the random variables w and d, the number of latent topics is usually equal with the expected number of object categories in the data set. The joint distribution can be decomposed into two simpler conditionals and a document prior.

$$P(w,d) = \sum_{z \in Z} P(w,d,z)$$

$$= \sum_{z \in Z} P(w \mid z) P(z \mid d) P(d) \qquad (2.2)$$

The expectation maximization algorithm EM [71] is used to learn in an unsupervised manner the conditionals probabilities. The new documents d are classified by running EM with $P(w \mid z)$ fixed. As a result $P(z \mid d)$ is obtained this is a mix of topics in that document. Even if it was refer to words and documents, the co-occurrence of any couple of discrete variables may be modeled in exactly the same way. Great pLSA advantages on the modeling side are: the well defined probabilities, the interpretable directions in the Probabilistic Latent Semantic space as multinomial word distributions and very important, a better selection of the model and complexity control

In [63] several LSA methods have been applied to visual recognition and in [72] Fergus have incorporated spatial information into a pLSA.

In [73] Blei proposes the latent dirichlet allocation. Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus, where a corpus is a collection of documents. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. This is similar to probabilistic latent semantic analysis (pLSA), except that in LDA the topic distribution is assumed to have a Dirichlet prior.

According to [74] LDA is a MAP / ML estimated LDA model under a uniform Dirichlet distribution. In [11] is shown an application of LDA to scene analysis

2.1.2 Geometry and Appearance

Another solution are part-based models which are a popular choice for enriching view-based approaches with global object geometry. Illustration of different part-based models can be seen in figures below. These models are build on the original idea of Fischler and Elschlager [75] of using the relative position of a few template matches and evolved in complexity in the work of Perona and others. There are several models in literature proposed for combining local appearance in a common object representation. In figures below different types of part-based models can be seen. Part-based models have recently applied to recognition problems with many categories (more than a hundred) and large intra-class variations. The simplest part-based models, bag-of-feature, without spatial structure have become really popular in the last years.

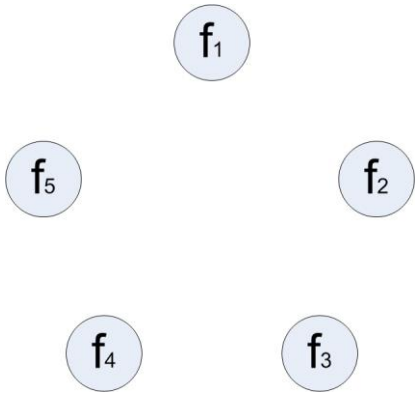


Figure 2 Bag-of-features: all parts are assumed to be independent.

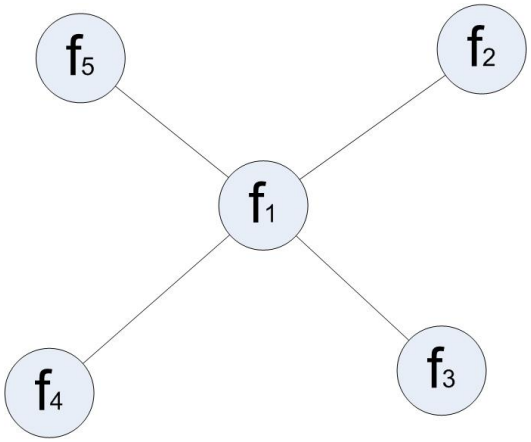


Figure 3 Star graph: there exists one reference part (f1) on which all other parts are conditioned

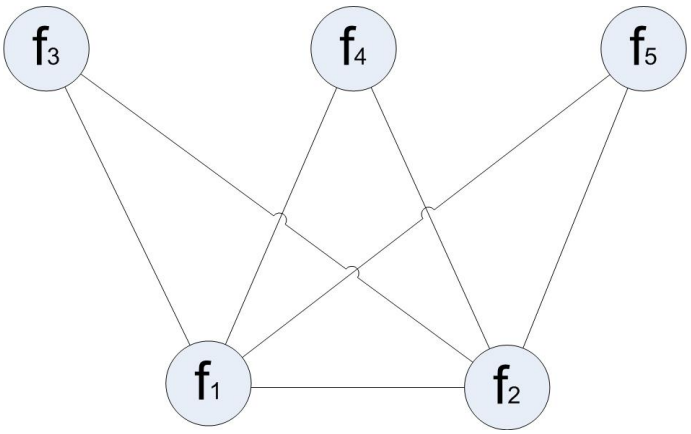


Figure 4 k-fan (k = 2): k reference parts on which all other parts are conditioned (f1 and f2 in this example).

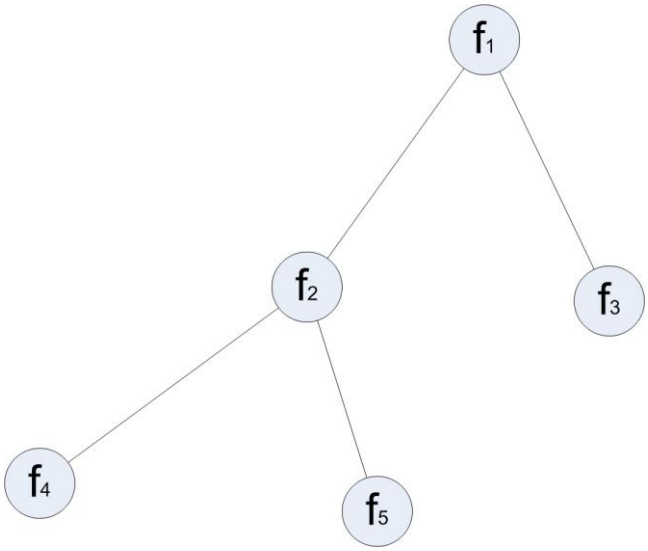


Figure 5 Tree model: parts dependencies form a tree hierarchy

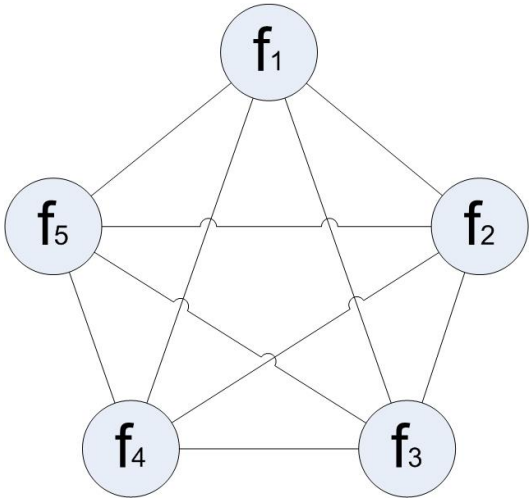


Figure 6 Constellation model: each part is depending on all other parts resulting in a fully connected graph.

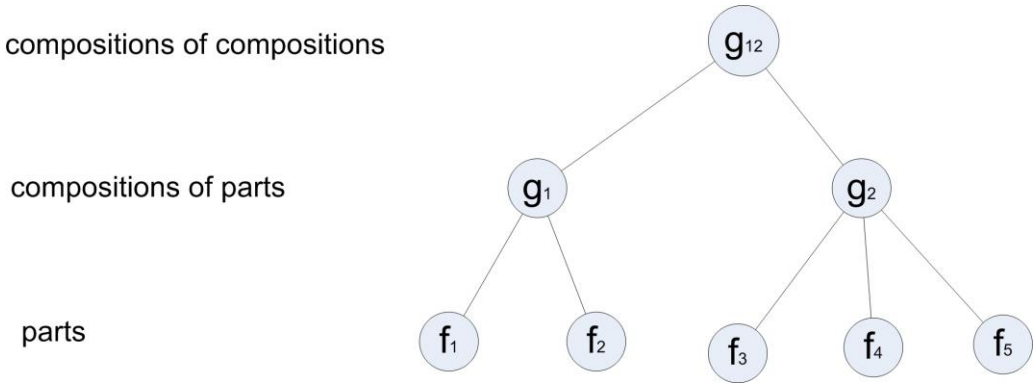


Figure 7 Compositional hierarchy: intermediate compositions of parts, the g_i, are established. In contrast to tree model, parts are only present at the leafs of the hierarchy. The intermediate compositions g_i are not observed directly, but inferred from the parts.

2.1.2.1 Parts and Structure Models

For the first time a model that combined local parts and global spatial structure was proposed in [75]. This model is referred to as parts and structure model, the model consists of a series of small templates (the parts) arranged in some geometric configuration (the structure).
Following the same idea and inspired by the dynamic link architecture for cognition process in [76] a recognition system based on deformable grid template is proposed. In [77] edge fragments are grouped than exhibit stable relative positions. The groupings are agglomerated into a joint object representation and during training there are sough that discriminate one object class from the rest.

2.1.2.2 Biologically Inspired Convolutionar Networks

One of the first influential biological inspired model is the one described in [78]. A neural network model for a mechanism of visual pattern recognition is proposed in this paper. The network is self-organized by "learning without a teacher", and acquires an ability to recognize stimulus patterns based on the geometrical similarity (Gestalt) of their shapes without affected by their positions. This network is called by the author "neocognitron". The neocognitron is inspired from the model proposed by Hubel and Wiesel [60]. They found two types of cells in visual primary cortex called simple cell and complex cell, alternating layers of S and C types neurons constitute the feed-forward Neocognition. The S-neurons show characteristics similar to simple cells or lower order hypercomplex cells, and the C-neurons similar to complex cells or higher order hypercomplex cells. The afferent synapses to each S-cell have plasticity and are modifiable. After repetitive presentation of a set of stimulus patterns, each stimulus pattern has become to elicit an output only from one of the C-cells of the last layer, and conversely, this C-cell has become selectively responsive only to that stimulus pattern. That is, none of the C-cells of the last layer responds to more than one stimulus pattern. The response of the C-cells of the last layer is not affected by the pattern's position at

all. Neither is it affected by a small change in shape nor in size of the stimulus pattern. In other words S-cells build up feature complexity while C-cells introduce location invariance. This system has been used for handwritten character recognition and other pattern recognition tasks.

In [79] is proposed a large neural networks that receive pixels from small images as input and return the class label as output. The layers from the convolutional neural network alternate between sub-sampling and convolution operations (S-cell and C-cell). Their five-layer convolutional network is somewhat similar to LeNet-5, but with multiple input planes and different numbers of units on the last two layers. Neurons in the output layer compute the distance of their input to a pattern that is stored in the weights. The network is applied to the problem of handwritten digit recognition. The spatial extend of the receptive fields of neurons increases towards the output layer, so the occluded or corrupted parts of an image have a significantly lower influence on neurons in higher layers than on those of the input layer. During training the network weights are learned using back-propagation on labeled training digits. Due to the large number of weights large training sets are required.

HMAX [80] model is another biologically inspired model. Like other convolutional networks this model consists of alternating S and C layers. The input layer consists of a Gabor filterbank [81] and , the sub-sampling is performed by computing the maximum over all inputs of a C-cell rather than the weighted sum. HMAX is a multi-scale model since C-cells pool information not only over locations but also over nearby scales. Later extensions are [82], [83], [84].

2.1.2.3 Constellation models

Among part-based models, a very popular one seems to be the constellation model, which attempts to represent an object class by a set of N parts under mutual geometric constraints. The constellation model was developed by Perona and his team [85], [86], [87] and its representing a probabilistic adaptation of Fischler and Elschlager [75] approach. Constellation model differs significantly "bag-of-words" representation models, which explicitly disregard the location of image features. Constellation Model is used that explicitly handles missing features and background clutter, in addition to representing the spatial layout of the parts in the model. A particular attraction is that the authors designed a system which requires minimal supervision to train, even on cluttered images.

The model of Fischler and Elschlager was revisited by Burl et al.[88], [85]. Natural images of faces were the input to the system. A face model was manually trained by identifying fiducial points on a set of training faces, giving statistics for a set of detectors as well as joint statistics of their relative location.

Weber et al.[89], [90], [86] improved Burl et al. approach by training the model using a more unsupervised learning process, which precluded the necessity for tedious hand-labeling of parts. Their algorithm was particularly remarkable because it performed well even on cluttered and occluded image data. This was done by automatically obtaining a set of potentially useful pixel patches by running an interest operator on the training set; chopping out patches around each interest point and then using k-means clustering on the patches. The cluster centres are used as detectors to provide a set of points from which the shape model is learnt. The detections from various combinations of cluster centres are used in turn to build a set of shape models. Each model was trained on a training set using EM and

tested on a separate validation set. The final model selected was the one which gave the best performance on the validation set.

Fergus et al.[87] improved later this model. They used a probabilistic framework and make the learning step fully unsupervised, having both shape and appearance learned simultaneously, and accounting explicitly for the relative scale of parts.

In [91] Fergus et al. estimate the joint Gaussian distribution of spatial arrangement, scale, appearance, and edge curves in all detected patches. The number of parameters in this model grows exponentially with the number of parts and, therefore, the complexity of the joint model causes only small numbers of parts to be feasible. Consequently, the approach is only suitable for object classes that can be characterized with very few, highly specific parts. An object representation that is founded on very few, but highly specific components suffers from the problem that such critical parts can only be detected with limited reliability

Fei-Fei et al. [32], [92] give an example of the application of powerful machine learning methods to the Constellation Model. They introduce a hierarchical Bayesian version of the Constellation Model which is able to incorporate priors into the learning procedure in a principled manner. This enables the algorithm to train from very few images (< 5) rather than the hundreds typically required.

2.1.2.4 Models with Large Numbers of Appearance Patches

As it was said above the constellation models usually have a small numbers of parts. In papers like [93], [94], [95], [96] models which are based on a large number of local image patches are investigated. In the manner of Weber et al., Agarwal and Roth [93], [94] extract appearance patches at interest points and cluster them to find common patches. Vectors are formed from combinations of features in the image and their spatial relations are encoded in a coarse manner. An image is represented by all patch prototypes that have been detected and a set of spatial relations between them. Their algorithm uses a window of interest which is moved exhaustively over all locations and scales in an image to obtain candidate object hypotheses. Hypotheses that cover background are then filtered out with a sparse network of windows (SNoW) classifier [97].

Borenstein and Ullman [98] present a scheme which combines object classification with segmentation. The object is modeled by a small set of image fragments; small rectangular textured patches, distinctive of the class. In recognition, full coverage of the object is obtained by combining the fragments, since each fragment is accompanied by foreground-background mask, a pixel-level segmentation of the test image may be obtained.

Thereafter Leibe and Schiele [96] improved Agarwal and Roth approach with a star graph like shape representation. Appearance patches are clustered to obtain prototypical representatives of object parts but additionally the relative location of the object center is recorded for each part during training and also a foreground/background mask is stored. This approach has two main advantages: it allows a validation step which measures the support, on a pixel level, of the fit between the model and the image, and permits a pixel level segmentation of object from test instances in the manner of [98]. In recognition interest points are again found and then a probabilistic Hough scheme used to vote for the position of the object within the image, based on the match of the regions to each of the clusters.

The maximum in voting space is then selected as the position of the object center and it is used to discard irrelevant parts in a subsequent back projection stage. The foreground/background masks which are associated with each remaining prototypical part are averaged to segment the object from the background. Leibe and Schiele approaches' compared to constellation models require more supervision during training; therefore the approach is limited to small numbers of object categories. In [95] the approach is extended to deal with objects on multiple scales. This method usually is applied to discriminate one category from background, but lately was applied to pedestrian recognition [99], [100].

2.1.2.5 Tree and k-fan Models

Felzenszwalb and Huttenlocher [101] have proposed part-based models where the spatial relationships between parts are tree-structured for computational reasons. The application for their model was to find people in images. The human being articulated structure was represented as a tree-structure (Figure 5), the main contribution of this work is the complexity reduction of the matching algorithm.

In [102] Cradall et al. investigate part models of different complexity. In this paper is introduced a class of graphs that is called k-fans (see Figure 4). Graphical models defined by k-fans provide a natural family of spatial priors for part-based recognition, where k specifies the complexity of the model. A 0-fan is a bag-of-features model with no dependencies, star graph models are 1-fans, and a fully connected constellation model of N parts corresponds to an N-fan. Their conclusion was that relatively simple models provide similar performance to complex fully connected spatial models, but have a substantially lower computational cost.

2.1.2.6 Shape Matching for Solving the Correspondence Problem

Berg and Malik [47] propose a scheme based on deformable shape matching where the correspondence between the model and features in the image is posed as an integer quadratic programming problem.Their formulation is able to deal with outliers when estimating the correspondence, thus occlusion and background clutter can be handled. Once correspondence has been estimated, a thin-plate spline is used for give an aligning transform between the model and the test exemplar, giving a dense correspondence between the two.

The nearest neighbor approach [42], [103] that matches a query against all training samples for all categories conducted to the correct object class. Combining the nearest neighbor classifier with a support vector machine (SVM) [104] better classification rates were obtained [105].

2.1.2.7 3-D Geometric Context

Some papers that involved the information on the 3-D nature of object are [106], [107], [108] and in others [109] scene context was used as source of information. In some situations, contextual information can provide more relevant information for the recognition of an object than the intrinsic object information, so the idea was to detect objects exclusively based on contextual cues by captureing those characteristics of a scene that are indicative for the presence of object categories.

2.2 Hierarchical Object Models and Compositionality

During the years the researchers from computer vision field tried to fill the semantic gap between the low-level features and the abstract nature of the models. The earlier systems rather than to address this representational gap eliminated it by bringing the images closer to the models, by removing object surface markings and structural detail, controlling lighting conditions, and reducing scene clutter. Edges in the image were assumed to map directly to the limbs and surface discontinuities of high-order volumetric parts making up the models. Thereafter, the representational gap was eliminated by bringing the model closer to the imaged object, requiring the model to capture the exact geometry of the object. When the initial measured representation and the final concept representation are far away from each other, a direct modeling requires complex models that are cumbersome and difficult to learn. The computational complexity was seriously affected buy the presence of texture and surface markings, so objects which were texture free were preferred. The resulting systems were unable to recognize complex objects with complex surface markings.

Most of these earlier model-based approaches were based on shallow hierarchies. In this context, hierarchies have been employed mainly because model based representations are rather abstract and high level in the sense that they are not focused on the observed images but on the physical, 3-D objects being imaged. Building such abstract concepts from a concrete image requires a complex abstraction process. To simplify this process, it has been divided into several smaller steps resulting in intermediate representations. A classical example is Marr's [110] primal, 2.5D sketch and 3D model.

Once the view-based models have replaced CAD models, representational gap was eliminated by bringing the models all the way down to the image.

View-based approaches perform classification as directly as possible in the image space so these are typically non-hierarchical. Simple view-based models that incorporate very limited geometry or even no spatial structure at all can be represented directly without intermediate stages. Examples are template matching, eigenfaces, global image histogram, or simple bag-of-features approaches. If the spatial structure is wanted to be incorporated into a representation, the complexity rises rapidly with the number of model parts. This can be seen in the case of constellation models.

When the initial measured representation and the final concept representation are far away from each other, a direct modeling is not a proper solution because learning object models corresponds to learning a mapping from the initial measurements into the space of abstract object representations. When both layers are too far apart it is favorable to replace the complex mapping by a concatenation of simpler ones that correspond to individual layers of a hierarchy. It is important to know that the number of layers cannot be arbitrary high because noise and other disturbances at the feature level can be amplified by successive representation layers.

There are several view-based approaches which have presented shallow hierarchies, for instance the bag-of-features approach, which is suited for large numbers of parts, has been extended by pLSA and LDA which introduce a single hidden layer [63]. On the other end of the modeling spectrum are the heavily supervised models of Felzenszwalb and Huttenlocher [101]. These are tree-

structured representations of only few parts that are modeled by hand. Other examples are[111] ,[112].

Though the appearance-based methods and machine learning algorithms have made remarkable progress, they have intrinsic problems that could be complemented by structure based methods.

The recent vision literature has observed a pleasing trend for returning to the compositional and grammatical methods, for example, the work in the groups of Ahuja [113], Geman [114], [115], Dickinson [116], Pollak [117], Buhmann [10] and Zhu [118], [119]. The return to these methods is in response to the limitations of the appearance based and machine learning methods when they are scaled up and is powered by progresses in several aspects like:

- consistent mathematical and statistical framework to integrate various image models, such as Markov (graphical) models, sparse coding [120], and stochastic context free grammar [121];
- more realistic appearance models for the image primitives to connect the symbols to pixels
- more powerful algorithms including discriminative classification and generative methods, such as the Data-Driven Markov China Monte Carlo [122] and a huge number of realistic training and testing images.

2.2.1 The Origin of Research on Compositionality

Compositionality has been studied in many diverse fields such as linguistics, logic, and neuroscience [123], but it is especially evident in the syntax and semantics of language. A limited number of letters (atomic constituents) can form a huge variety of words and sentence. The sentences are used to describe an unimaginable number of different scenarios. Although Gottlob Frege has highlighted the separability of sentences and thoughts into sub-structures in [124], probably the first contribution that has actually introduced the word "compositionality" is the paper of Katz and Fodor [125].

Werning et al. [123] propose a concise formulation that serves as a good approximation to this principle:

"An interpreted representational system R is compositional if and only if for every complex representation r of R, the meaning of r is determined by the structure of r and the meaning of the constituents of r."

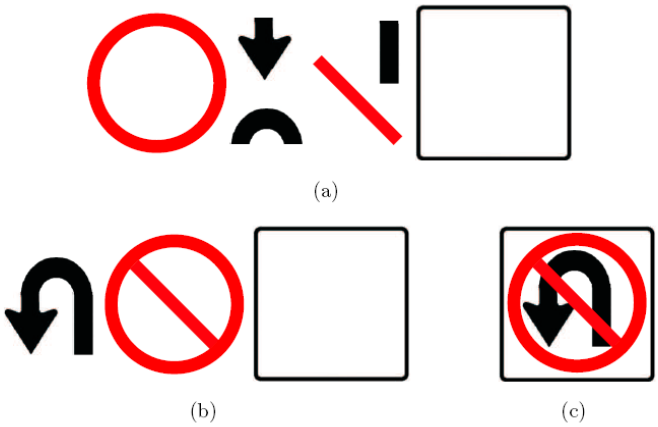


Figure 8 Illustration of compositionality according to [126]. a) Some generic geometric primitives. b) An intermediate, compositional grouping that highlights already some spatial relations. c) The simple generic parts together with spatial relations between them can be used to represent various objects (here a no U-turn traffic sign) although the parts themselves provide only very little information about the object.

In linguistics structure is defined by orthographic, syntactic, and grammatical rules, the primitive and atomic constituents are letters in written language, and phonemes in spoken language; this serve as fix points for recursive decomposition.

Biederman [33] gives an analysis of compositionality in vision: a restricted number of atomic shape primitives, called geons, constitute the components of more complex objects and a hierarchical construction process of entities can serve as a basis for recognition. The detection of an object is based on the spatial relationships (the structure) of its components.

2.3 Vision based hand gesture recognition

Hand gestures are a powerful, and probably one of the most natural and intuitive, human to human communication modality. Gesture recognition is nowadays an active topic of vision research which has applications in diverse fields such as: interactive games, performance analysis, surveillance monitoring, and remote control of home appliances, virtual reality, disability support, medical systems, sign language translation and many others.

In the last years there has been a great emphasis in HCI research to create interfaces that directly employ the natural communication and manipulation skills of humans. Gesture languages made up of hand postures (i.e., static gestures) or motion patterns (i.e., dynamic gestures[127], [128]) have been employed to implement command and control interfaces [129], [130], [131].

The main difficulties that show up when a hand pose estimation system is designed are:

High-dimensional problem: The hand is an articulated object with more than 20 DOF.

Even if not all 20 DOF are used to model the hand, studies have shown that it is not possible to use less than six dimensions; beside this also the location and orientation of the hand itself must be estimated so , there still exist a large number of parameters.

Self-occlusions: Since the hand is an articulated object, its projection results in a large variety of shapes with many self-occlusions, making it difficult to segment different parts of the hand and extract high level features.

Processing speed: Even for a single image sequence, a real-time CV system needs to process a huge amount of data. With the current hardware technology, some existing algorithms require expensive, dedicated hardware, and possibly parallel processing capabilities to operate in real-time.

Uncontrolled environments: Almost all HCI system are expected to operate without background restrictions and a wide range of lighting conditions, on the other hand, even locating a rigid object in an arbitrary background is almost always a challenging issue in computer vision.

Rapid hand motion: The hand has very fast motion capabilities with a speed reaching up to 5 m/s for translation and 300degree/s for wrist rotation.

The approaches to Vision based hand posture and gesture recognition can be divided into two categories – 3 D hand model based approaches and appearance based approaches.

2.3.1 3 D hand model based approaches

Model based approaches attempt to infer the pose of the palm and the joint angles, such an approach would be ideal for realistic interactions in virtual environments. Generally, the approach consists of searching for the kinematic parameters that brings the 2D projection of a 3D model of hand into correspondence with an edge-based image of a hand.

A 27 DOF model was introduced in [132] and has been used in many studies an it is shown in Figure 10 b. The CMC joints are assumed to be fixed, which quite unrealistically models the palm as a rigid body. The fingers are modeled as planar serial kinematic chains attached to the palm at anchor points located at MCP joints. The planarity assumption does not hold in general. Standard robotics techniques provide efficient representations and fast algorithms for various calculations related to the kinematics or dynamics of the model. Adding an extra twist motion to MCP joints [133], [134] introducing one flexion/extension DOF to CMC joints [135] or using a spherical joint for TM [136] are some examples of the variations of the kinematic model.

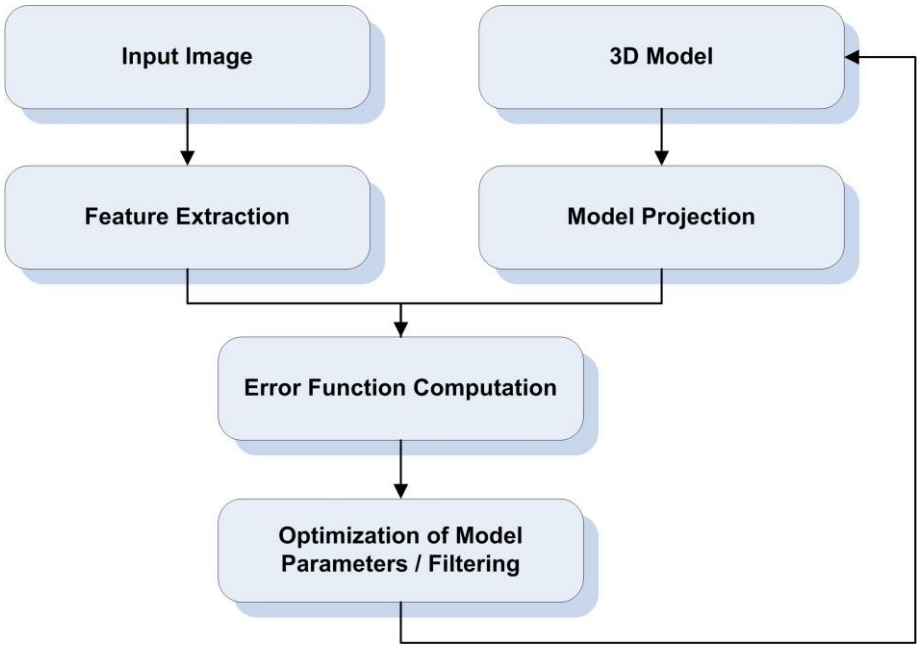
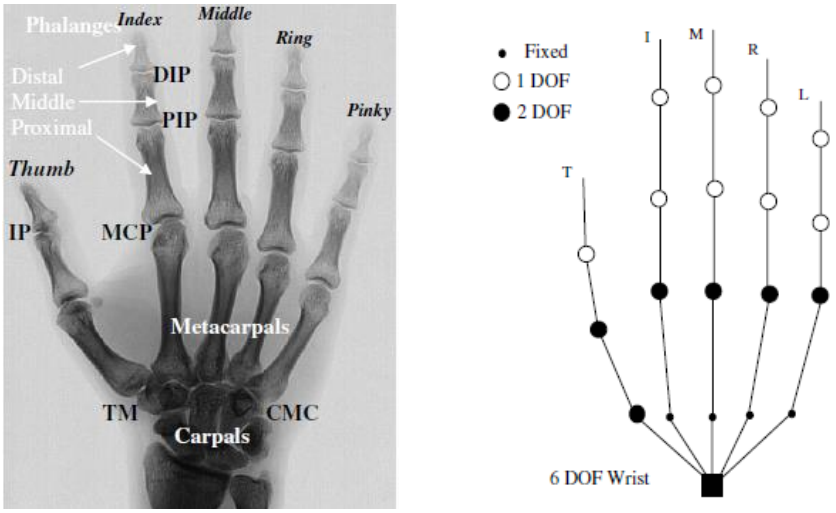


Figure 9 Components of a model-based tracking system. A model-based tracking system employs a geometric 3D model, which is projected into the image. An error function between image features and model projection is minimized, and the model parameters are adapted.

One of the earliest model based approaches to the problem of bare hand tracking was proposed by Rehg and Kanade [137]. This article describes a model-based hand tracking system, called DigitEyes, which can recover the state of a 27 DOF hand model from ordinary gray scale images at speeds of up to 10 Hz. The hand tracking problem is posed as an inverse problem: given an image frame (edge map) find the underlying parameters of the model. The inverse mapping is non-linear due to the trigonometric functions modeling the joint movements and the perspective image projection. A key observation is that the resulting image changes smoothly as the parameters are changed. Therefore, this problem is a promising candidate for assuming local linearity. Several iterative methods that assume local linearity exist for solving non-linear equations (e.g. Newton's method). Upon finding the solution for a frame the parameters are used as the initial parameters for the next frame and the fitting procedure is repeated. The approach can be thought of as a series of hypotheses and tests, where a hypothesis of model parameters at each step is generated in the direction of the parameter space (from the previous hypothesis) achieving the greatest decrease in miscorrespondence. These model parameters are then tested against the image. This approach has several disadvantages that has kept it from real-world use. First, at each frame the initial parameters have to be close to the solution, otherwise the approach is liable to find a suboptimal solution (i.e. local minima). Secondly, the fitting process is also sensitive to noise (e.g. lens aberrations, sensor noise) in the imaging process. Finally, the approach cannot handle the inevitable self-occlusion of the hand.



a) b)

Figure 10 Skeletal hand model: (a) Hand anatomy, (b) the kinematic model according to [138]

Heap et al.[139] proposed a deformable 3D hand model and modeled the entire surface of the hand by a surface mash constructed via PCA from training examples. Real-time tracking is achieved by finding the closest possibly deformed model matching the image. Such a representation requires further processing to extract useful higher-level information, such as pointing direction; however, it was shown to be very effective to reliably locate and track the hand in images. The method however, is not able to handle the occlusion problem and is not scale and rotation invariant.

Figure 11 Hand tracking using 3D Point Distribution Model from [139]

In [140] is proposed a model called cardboard model which is able to capture articulated hand motions. The constrains on the joint configurations are learned from natural hand motions using a data glove as input device, then a

sequential Monte Carlo tracking algorithm, based on importance sampling, produces good results, but is view-dependent, and does not handle global hand motion.

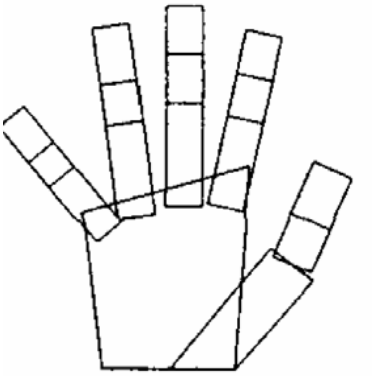


Figure 12 Cardboard model taken from [140]

Stenger et al. [141] used quadrics as shape primitives, as shown in FIG. This allows for the generation of 2D profiles of the model using elegant tools from projective geometry, and for an efficient method to handle self-occlusion. The pose of the hand model is estimated with an Unscented Kalman filter (UKF) [142], which minimizes the geometric error between the profiles and edges extracted from the images. The use of the UKF permits higher frame rates than more sophisticated estimation methods such as particle filtering, whilst providing higher accuracy than the extended Kalman filter. More recent efforts have reformulated the problem within a Bayesian (probabilistic) framework [143]. Bayesian approaches allow for the pooling of multiple sources of information (e.g. system dynamics, prior observations) to arrive at both an optimal estimate of the parameters and a probability distribution of the parameter space to guide future search for parameters. On contrary to Kalman filter approach, Bayesian approaches allow non-linear system formulations and non- Gaussian (multi-modal) uncertainty (e.g. caused by occlusions) at the expense of a closed-form solution of the uncertainty. A potential problem with the approach is that certain independent assumptions of the underlying probabilistic distribution are made, for computational tractability reasons that may not hold in practice. Also, it is a computationally expensive approach.

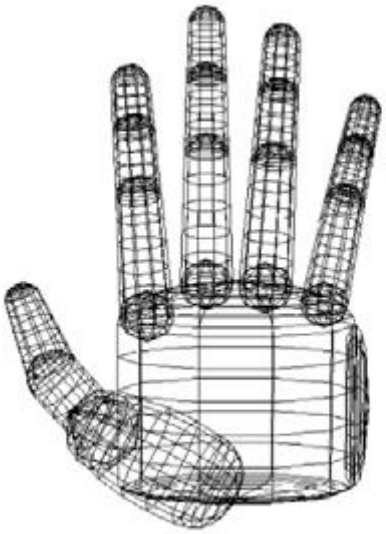


Figure 13 Quadrics-based hand model taken from [141]

Three dimensional hand model based approaches offer a rich description that potentially allows a wide class of hand gestures. However, as the 3D hand models are articulated deformable objects with many DOF's, a very large image database is required to cover all the characteristic shapes under different views. Another common problem with model based approaches is the problem of feature extraction and lack of capability to deal with singularities that arise from ambiguous views.

2.3.2 Appearance-Based Models

Appearance-based models are derived directly from the information contained in the images and have traditionally been used for gesture recognition. This is often posed as a pattern recognition problem, which may be partitioned into components such as shown in Figure 14

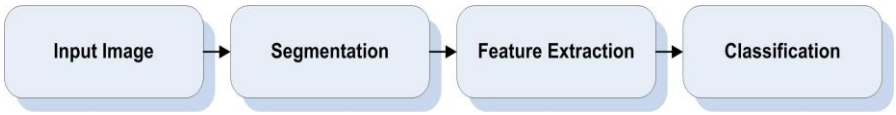


Figure 14 Components of a view-based recognition system. View-based recognition is often treated as a pattern recognition problem and a typical system consists of components as shown above. At the segmentation stage the hand is separated from the background. Features are then extracted, and based on these measurements the input is classified as one of many possible hand poses. The classifier is designed using a set of training patterns.

These methods are often described as bottom-up methods, because low-level features are used to infer higher level information. The main problems, which need to be solved in these systems, are how to segment a hand from a general background scene and which features to extract from the segmented region. The classification itself is mostly done using a nearest neighbor classifier or other standard classification methods [42]. Generally, the classifier is learnt from a set of training examples and assigns the input image to one of the possible categories

2.3.2.1 Hand localization

A straightforward and simple approach that is often utilized is to look for skin colored regions in the image. This is a very popular method [144], [145], [146], [147] but has some drawbacks. First, skin color detection is very sensitive to lighting conditions. While practicable and efficient methods exist for skin color detection under controlled (and known) illumination, the problem of learning a flexible skin model and adapting it over time is challenging.

Static background subtraction [148], [149], and the use of adaptive background models [150],[151] are also common methods.

Shadows can be a problem in background subtraction algorithms [151]. A few studies utilize IR cameras that are tuned to human temperature [152], [153] to provide fast solutions by simple threshold operations. Various assumptions are used, such as the hand being the only skin-colored object, uniform ambient lighting, or stationary background.

According to [154] motion cue is also commonly applied for gesturer localization and is used in conjunction with certain assumptions about the gesturer. The main component of motion in the visual image is usually the motion of the hand of the gesturer and can thus be used to localize it. This localization approach is used in [155], [156]. The disadvantage of the motion cue approach is in its assumptions. While the assumptions hold over a wide spectrum of cases, there are occasions when more than one gesturer is active at a time or the background is not stationary.

Another approach to locate the hand involves classification-based object detection methods. In these systems, a large list of hypotheses in the form of

subregions in the image is processed by a classifier to decide the presence of the object. In [157] is introduced a class separability estimation method based on frequency spectrum analysis to reduce the load of training these classifiers. Ong et al.[158] employed clustering methods to cluster the training data into similar shapes and build a tree of classifiers for detection. As one goes down to the branches of the tree, the classifiers are trained to detect more and more specific clusters consisting of similar shapes, this classifier can be used to recognize gestures as well. Wu et al.[159] provided a training algorithm that only needs a small portion of the database to be labeled and built a viewin dependent gesture recognition system by using samples of postures captured from a large number of views. A color segmentation algorithm [99] was employed to reduce the number of hypotheses.

2.3.2.2 Feature extraction

Hand silhouette is among the simplest, yet most frequently used features. Silhouettes are easily extracted from local hand and arm images in the restrictive background setups. By silhouette, it is meaning the outline of the hand provided by segmentation algorithms, or equivalently the partitioning of the input image into object and background pixels. In the case of complex backgrounds, techniques that employ color histogram analyses can be used. In order to estimate the hand pose the position and orientation of the hand, fingertip locations, and finger orientation from the images is extracted. The center of gravity (or centroid) of the silhouette [160], [161] is one choice, but it may not be very stable relative to the silhouette shape due to its dependence on the finger positions. The point having the maximum distance to the closest boundary edge [144], [152], [162], [163] as been argued to be more stable under changes in silhouette shape.

A frequently used feature in gesture analysis is the fingertip. Fingertip detection can be handled by correlation techniques using a circular mask [152], [153], [151], which provides rotation invariance, or fingertip templates extracted from real images [164], [165]. Another common method to detect the fingertips and the palm-finger intersections [148], [149], [147] is to use the curvature local maxima on the boundary of the silhouette. For the curvature-based methods, in case of noisy silhouette contours, the sensitivity to noise can be an issue. In [144], [161], a more reliable algorithm based on the distance of the contour points to the hand position is utilized. The local maximum of the distance between the hand position and farthest boundary point at each direction gives the fingertip locations. The direction of the principal axis of the silhouettes [166], [163], [148] can be used to estimate the finger or 2D hand orientation. All these features can be tracked across frames to increase computation speed and robustness using Kalman filters [144], [167] or heuristics that determine search windows in the image [147], [168] based on previous feature locations or rough planar hand models. The low computational complexity of these methods enables real-time implementations using conventional hardware but their accuracy and robustness are arguable. These methods rely on high quality segmentation lowering their chance of being applied on highly cluttered backgrounds. Failures can be expected in some cases such as two fingers touching each other or out of plane hand rotations.

Jennings et al.[169] have used a more elaborate method by tracking the features directly in 3D using 3D models; their method has employed range images, color, and edge features extracted from multiple cameras to track the index finger in a pointing gesture. Very robust tracking results over cluttered and moving

backgrounds were obtained. Cylindrical fingertip models to track multiple finger positions in 3D without occlusions, over uniform backgrounds were used by Davis et al. [170]. Markers could also be a solution, although they are considered intrusive they have considerable technical advantages in terms of processing speed. In [168] elliptical markers to estimate the hand frame in 3D re used; and in [171] white fingertip markers under black-light to detect fingertip locations were used, yielding a much richer set of gestures.

View-based methods have been shown to be effective at discriminating between a certain number of hand poses, which is satisfactory for a number of applications in gesture recognition. One of the main problems in view-based methods is the segmentation stage. This is often done by skin colour segmentation, which requires the user to wear long sleeves or a wristband. Another option is to test several possible segmentations, which increase the recognition time.

3 THEORETICAL BACKGROUNDS

3.1 Introduction

In this chapter some theoretical fundaments from rather different fields are presented. These theoretical fundaments are used in chapter four to implement the hand poster recognition system.

3.2 Canny edge detector

An important subproblem in computer vision is how to detect edges from grey-level images. Edges represent commonly used features. The importance of edge information for early machine vision is usually motivated from the observation that under rather general assumptions about the image formation process, a discontinuity in image brightness can be assumed to correspond to a discontinuity in depth, surface orientation, reflectance or illumination. In this respect, edges in the image domain constitute a strong link to physical properties of the world.

In this work contours have an important role as it can be seen in the next chapter. It is easy to understand that a large number of approaches have been developed for detecting edges. The earliest schemes focused on the detection of points at which the gradient magnitude is high. Derivative approximations were computed either from the pixels directly, using operators such as Robert's cross operator [172], the Sobel operator [173] and the Prewitt operator [174], or from local least-squares fitting [175]. All the gradient-based algorithms have kernel operators that calculate the strength of the slope in directions which are orthogonal to each other, commonly vertical and horizontal. Later, the contributions of the different components of the slopes are combined to give the total value of the edge strength. Gradient-based algorithms such as the Prewitt filter have a major drawback of being very sensitive to noise. The size of the kernel filter and coefficients are fixed and cannot be adapted to a given image. An adaptive edge-detection algorithm is necessary to provide a robust solution that is adaptable to the varying noise levels of the images. Canny considered the problem of determining an "optimal smoothing filter" of finite support for detecting step edges.

In order to detect the hand contours in this work John Canny's approach to optimal edge detection was fallowed. In this section this approach is reviewed.

In his work [176], [177], Canny specifies three performance criteria that an edge detector should fulfill. The criteria Canny uses to detect step edges can be summarized as follows:

Good detection: Edge detection should be robust to noise; it means that should be a high probability to mark real edge points and a low probability to falsely mark non edge points. This criterion can be formalized [177], [178] by requiring the edge detector to maximize the output signal-to-noise ratio (SNR) for a given input signal-to-noise ratio.

Good localization: The points considered to be edge should be as close as possible to the center of the true edge. This criterion is formalized in [178] by requiring that the location of the maximal detector response to an edge exhibits low variance.

Uniqueness of response: The edge detector should produce only one response to a single edge. This criterion can be formalized by requiring the filter which performs edge detection to have a small spatial width [178].

The output signal-to-noise ratio can be increased by low pass filtering the input signal. This increase the spatial width of the overall filter that performs edge detection, as described in [178]. Therefore a trade-off between the first and third criterion has to be found. By using additional necessary properties of a filter that performs edge detection Tagare and deFigueiredo come to the conclusion that the derivative of the Gaussian—a Gaussian smoothing of the image followed by a first derivative—is the optimal detector.

An image consisting of three color channels for red, green, and blue, is denoted by I described by Eq.(3.1):

$$I(x,y) = (I_r(x,y), I_g(x,y), I_b(x,y))^T \tag{3.1}$$

I_G denotes the grayscale version of that image, i.e. its brightnesses. First the brightnesses I_G are smoothed by a (discrete) convolution—denoted by the $*$ operator—with a two-dimensional Gaussian G_c.

$$G_c(x,y) = \frac{1}{\sqrt{2\pi\sigma_c}} \exp\left(-\frac{x^2+y^2}{2\sigma_c{}^2}\right) \tag{3.2}$$

The strength of the smoothing is determined by σ_c, the standard deviation of the Gaussian filter. The smoothing filter used in the first stage directly affects the results of the Canny algorithm. Smaller filters cause less blurring, and allow detection of small, sharp lines. A larger filter causes more blurring, smearing out the value of a given pixel over a larger area of the image.

After the image has been convolved with a Gaussian, the edge direction is estimated from the gradient of the smoothed image intensity surface. The gradient of the smoothing, $\nabla(G_c * I_G)$, is computed on a discrete grid of image coordinates. The normal n_c to the target of the edge contour is estimated according to [177] by

$$\bar{\mathbf{n}}_c = \frac{n_c}{\|n_c\|_2} \tag{3.3}$$

where $$n_c = \nabla(G_c * I_G) \tag{3.4}$$

Potential edge points are local maxima of the gradient magnitude in the direction of the gradient, $\bar{n}_c$. The implementation of the resulting technique, which is called non-maximum suppression, examines the neighbors in gradient direction to both sides of a candidate edge point. This candidate point is only marked as an edge point if its gradient magnitude is greater than that of the neighbors.

The directional non-linear suppression is equivalent to the application of the following non-linear differential predicate:

$$\frac{\partial^2}{\partial \bar{n}_c^2} G_C * I_G = 0 \tag{3.5}$$

The neighbors surrounding the marked edge point are removed by setting their gradient magnitude to zero. The eight direct neighbors of each marked edge point that is not removed by

non-maximum suppression are examined in a postprocessing step using a lower and a higher threshold on gradient magnitude, t_{low} for high edge sensitivity and t_{high} for low edge sensitivity, respectively: Starting on an edge point whose magnitude of the gradient succeeds the higher threshold, all direct neighbors of this point are visited recursively. Each of these points is also marked as an edge point if its gradient has a magnitude that is greater than t_{low}. The described procedure is called hysteresis thresholding, and it continues recursively by examining the neighbors of all marked edge points. The use of two thresholds with hysteresis allows more flexibility than in a single-threshold approach. It is difficult to give a generic threshold that works well on all images. No tried and tested approach to this problem yet exists. The result is that small accidental gaps on edge contours are bridged.

3.3 The Harris Corner Detector

In computer vision many tasks rely on low-level features. The Harris corner detector is a popular interest point detector due to its strong invariance to [179]: rotation, scale, illumination variation and image noise. Harris corner detector is based on the local auto-correlation function of a signal; where the local auto-correlation function measures the local changes of the signal with patches shifted by a small amount in different directions.

Given a shift (u,v) and a point (x, y), the change of intensity for the shift is given in Eq. (3.6).

$$c(u,v) = \sum_{x,y} w(x,y)(I(x,y) - I(x+u,y+v))^2 \tag{3.6}$$

Window function Intensity Shifted Intensity

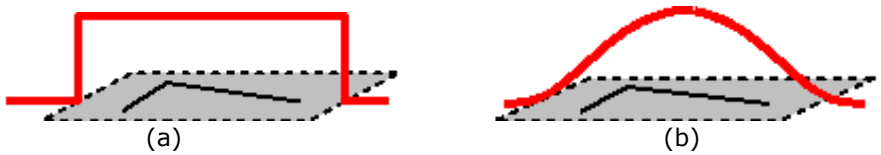


Figure 15 Window functions (a) 1 in window, 0 outside, or (b) Gaussian

The shifted image is approximated by a Taylor expansion truncated to the first order terms according to Eq.(3.7)

$$I(x+u,y+v) \approx I(x,y) + uI_x(x,y) + vI_y(x,y) \tag{3.7}$$

Substituting Eq.(3.7) in Eq.(3.6) the following result is obtained:

$$c(u,v) = \sum (u^2 I_x{}^2 + 2uvI_xI_y + v^2 I_v{}^2) \tag{3.8}$$

Rewritten as a matrix Eq.(3.8) yields to:

$$c(u,v) = \sum \begin{bmatrix} u & v \end{bmatrix} \begin{bmatrix} I_x{}^2 & I_xI_y \\ I_xI_y & I_y{}^2 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} u & v \end{bmatrix} \left(\sum \begin{bmatrix} I_x{}^2 & I_xI_y \\ I_xI_y & I_y{}^2 \end{bmatrix} \right) \begin{bmatrix} u \\ v \end{bmatrix} \tag{3.9}$$

For a small shift

$$c(u,v) \cong \begin{bmatrix} u & v \end{bmatrix} M \begin{bmatrix} u \\ v \end{bmatrix} \tag{3.10}$$

Where

$$M = \sum_{x,y} w(x,y) \begin{bmatrix} I_x{}^2 & I_xI_y \\ I_xI_y & I_y{}^2 \end{bmatrix} \tag{3.11}$$

where matrix M captures the intensity structure of the local neighborhood. c(u,v) is closely related to the autocorrelation function, with M describing its shape at the origin.

Let λ_1 and λ_2 be the eigenvalues of matrix M. The eigenvalues form a rotationally invariant description. There are three cases to be considered:

If both curvatures are small so the auto-correlation function is flat, then the windowed image region is approximately constant intensity (arbitrary shifts of the patch causes little change in c(u,v)).

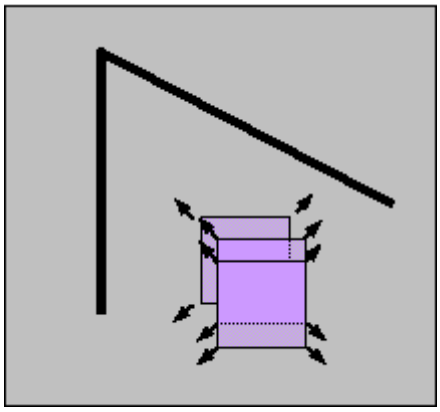


Figure 16 "Flat" region: no change in all directions

If one eigenvalue is high and the other low, so the local auto-correlation function is ridge shaped, then only local shifts in one direction (along the ridge) cause little change in c(u,v) and significant change in the orthogonal direction; this indicates an edge.

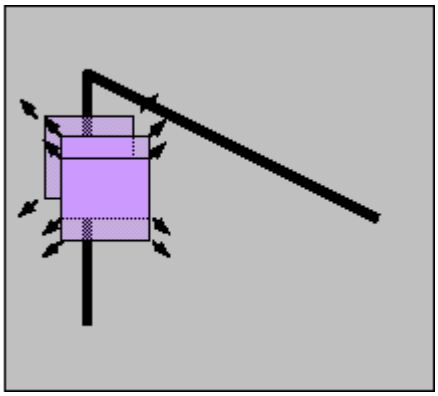


Figure 17 "Edge": no change along the edge direction

If both eigenvalues are high, so the local auto-correlation function is sharply peaked, then shifts in any direction will result in a significant increase; this indicates a corner.

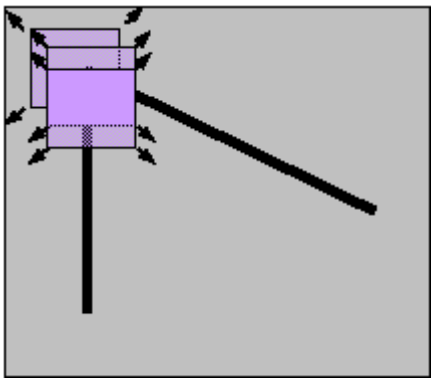


Figure 18 "Corner": significant change in all directions

Harris and Stephens note that exact computation of the eigenvalues is computationally expensive, since it requires the computation of a square root, and instead suggest the following function M_c, where κ a tunable sensitivity parameter is.

$$M_c = \lambda_1\lambda_2 - \kappa(\lambda_1 + \lambda_2)^2 = \det(A) - \kappa \mathrm{trace}^2(A) \qquad (3.12)$$

In literature there is a wide variety of interest points, such as: Harris [55], SUSAN (Smallest Univalue Segment Assimilating Nucleus) [180], SIFT (Scale-invariant feature transform) [181], DoG (Difference of Gaussian) [56], [181] etc. Comparative tests carried out amongst different approaches demonstrated that detectors which combine the smoothed image derivates via the autocorrelation matrix are robust and achieve very good results [179].

3.4 The Principle of Compositionality

According to [182] the visual stimulus is high redundant, it presents a significant spatial and temporal interdependency; the regularity makes portions of visual field to become predictable given other parts. Taking advantage of this dependency, compositionality is a general principle in cognition and especially in human vision [33]. Compositionality refers to the prominent ability of human cognition to represent entities as hierarchies of meaningful and generic parts. A representational system is compositional if and only if each complex representation is determined by its constituent parts and the relationships between them [123]. Compositional representations decompose complex objects into simpler parts, which are easier to recognize and using the relationships between them, finally resulting in a hierarchy of recursive compositions. It is based on a set of simple parts, like the Lego parts which can be used to build a large variety of objects. In compositionality a small number of generic low-level constituents is used to build an infinite number of hierarchically constructed entities in a different context. The Lego parts are not characteristic for any object that a child can build, there are quite simple and not so varied, but can be combining in a flexible way generating objects from houses to cars and robots to humans.

In order to create a new object to play with, the child dose not needs new parts (different shapes) to build it, the ones that already exits are used for the new scenario. The same is with compositionality: using a common set of low-level entities that are not characteristic for any category but generic, new scenarios of widely differing nature can be tackled without having to learn a novel low-level representation in order to adapt to new tasks. The object created by a child using a Lego is much more than its parts, comparative a compositional representation contains more information than what is merely present in its individual parts. A intuitive definition of the principle of compositionality is given by [183] in this sentences: "the whole is different from the sum of its parts". The flexibility of human vision that results from compositionality is also fundamental to computer vision, since it forms a basis for general applicability of a vision system. Consequently it can be summarized that compositions bridge the semantic gap between complex objects and the low level percepts (e.g. individual photoreceptive cells on the retina or, equivalently, pixels of a digital camera) by establishing intermediate hidden layer representations. Therefore, complex object models are decomposed into a hierarchy of simpler models so that learning those substructures becomes a feasible problem.

3.5 Gestalt psychology

Compositional representations decompose complex objects into simpler parts, which are easier to recognize and using the relationships between them, finally resulting in a hierarchy of recursive compositions; the Gestalt laws suggest how to organize the parts in order to form the compositions.

Gestalt psychology is a branch of psychology which has its origin in the work by Max Wertheimer [184]. The main idea is that the human visual system processes visual stimuli by grouping individual percepts to yield meaningful compositions. The underlying process of perceptual organization [29], [183] exhibits the principle law of Pragnanz, meaning that the "best, simplest, and most stable" groupings, [12] are preferred. Gestalt psychology has also proposed numerous Gestalt laws of organization [12] that are intended to entail Pragnanz. These laws represent composition rules that impose constraints on the types of admissible constructions, whereby most of the potential combinations are actually ruled out since they are rated meaningless.

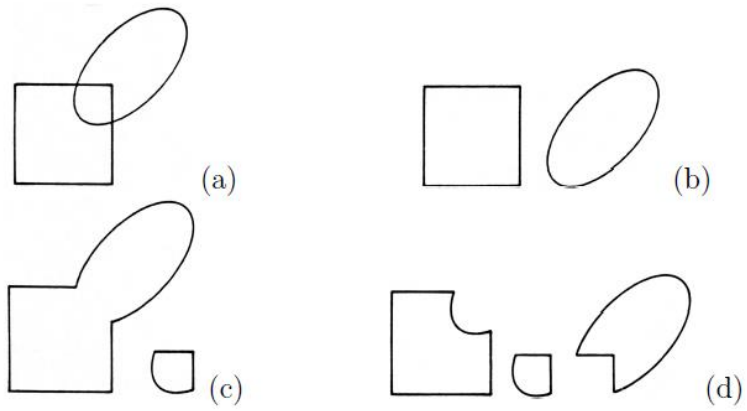


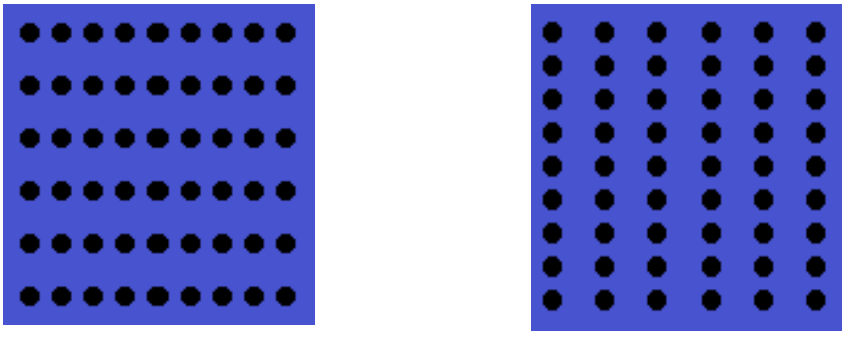
Figure 19 Law of Pragnanz: (a) Original image; (b), (c), and (d) are possible groupings; (b) is the grouping of greatest Pragnanz [183]

3.5.1 Gestalt laws

Gestalt psychology proposes a set of visual rules that guide the construction process of groupings and yield compositions of high Pragnanz. The Gestalt laws establish causal relationships between grouping constituents.

3.5.1.1 Proximity

One of the most important factors determining the perceptual organization of a scene is *proximity* of the elements within it. In Figure 20 a) and b) columns or rows dominate our focus; elements tend to be perceived as aggregated into groups if they are near each other.



(a) (b)

Figure 20 Illustration of the Gestalt laws: Proximity

3.5.1.2 Similarity

Similar features are associated; elements tend to be integrated into groups if they are similar to each other this can be seen in Figure 20, we tend to see alternating columns of circles and squares. The set of attribute that can be used to establish similarities includes:

- Orientation

Similarity of the spatial orientation of objects. The orientation is determined by prominent parts of an entity.

- Shape

Similarity of the form of objects. Figure 21 which is perceived as vertical columns, demonstrates this principle.

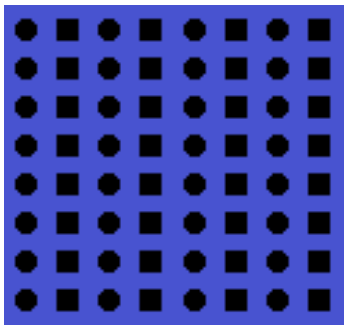


Figure 21 Illustration of Gestalt laws: Similarity

- Symmetry

Components that form a symmetrical composition tend to be grouped. Two contours that are similar when mirrored with respect to some axis are a common instance which this principle refers to. Figure 23 is perceived as three black patches since the resulting boundaries of each patch are symmetrical. The white area in between is observed as background. Figure 25 shows the same effects with switched colors.

- Color

Similarity in color of the objects or their direct surroundings. Color information about the local neighborhood is especially relevant when grouping boundary contours. Obviously, similarity in texture can be used in the same way.

- Common fate

Similarity of the motion pattern of objects: A camouflaged animal can be spotted much easier when it moves than while it remains stationary. Another example is the perception of a swarm of birds as a greater whole.

3.5.1.3 Good Continuity

In Figure 22 is more likely to identify lines a-b then a-c or a-d, and c-d then c-b or d-b. The lines which are based on smooth continuity are preferred over

abrupt changes, oriented units or groups tend to be integrated into perceptual wholes if they are aligned with each other

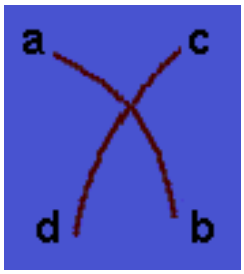


Figure 22 Illustration of Gestalt laws: Good Continuity

3.5.1.4 Closure

The mind may experience elements it does not perceive through sensation, in order to complete a regular figure (that is, to increase regularity); we tend to see three broken rectangles and a lonely box bracket] on the far left, rather than three girder profiles and a lonely box bracket] on right.

Figure 23 Illustration of Gestalt laws: Closure

3.5.1.5 Smallness

Smaller areas tend to be seen as figures against larger background. In Figure 24, we are likely to see a black cross rather than a white cross.

Figure 24 Illustration of Gestalt laws: Smallness

3.5.1.6 Symmetry

The principle of symmetry is that, the symmetrical areas tend to be seen as figures against the asymmetrical background (See Figure 25).

Figure 25 Illustration of Gestalt laws: Symetry

3.5.1.7 Surroundness

According to this principle areas which can be seen as surrounded by others tend to be perceived as figures. In Figure 26 the word is initially confused by resuming the black area as a ground. Once it assumes the background as surrounding area then you can discover the world 'Tie'

Figure 26 Illustration of Gestalt laws: Surroundness

3.6 Learning Paradigms

The goal of classification in general is to map a representation $x \in \Re^d$ of a real world object to an appropriate class label y ∈ {1, . . . , k}, where the representation x is already the result of a complex mapping that is a concatenation of a number of sub-processes such as image acquisition, preprocessing, feature extraction, and additional feature processing. According to [126] there are two ways of obtaining the optimal class label y: using *generative classifiers* which learn a model of the joint probability density function p(x,y) or equivalently of p(x|y) and p(y) and then the classification is perform based on Bayes formula; and *discriminative classifiers* which estimate the posterior p(y|x) directly, or by learning a direct mapping y=f(x) from inputs x to labels y.

An example of a visual object recognition system that learns from large volumes of data using a discriminative method is for instance the face classifier by Viola and Jones [40],[185]. Generative approaches are nevertheless popular [186], [96]. One reason for this popularity is that generative methods provide a straightforward way of dealing with missing or corrupted features by marginalizing the likelihood p(x|y) over the affected dimensions of x. In the context of limited training data, prior knowledge becomes increasingly important and a generative approach provides an intuitive solution to incorporating such a prior. Finally, an important property of generative models is that they allow to reverse the processing pipeline so that representations that are typical for learned object classes can be sampled and visualized. Almost all current approaches in object recognition are of probabilistic nature.

Bayes approach

While classifying some errors may be more expensive than others, for example, for a fatal disease that is easily cured by a cheap medicine without said effects false positives in diagnosis are better than false negatives. The loss function states exactly how costly each action is and is used to convert a probability determination into a decision.

Let $\{\omega_1,...,\omega_c\}$ be the set of *c* states of natures (categories) and let $\{\omega_1,...,\omega_a\}$ be the set of *a* possible actions. The loss function $\lambda(\omega_i \mid \omega_j)$ describes the loss incurred for taking action ω_i when state of nature is ω_j. The problem is to find a decision rule that minimize the overall risk. The overall risk *R* is the expected loss associated with a given decision rule.

$$R(\omega_i \mid x) = \sum_{j=1}^{c} \lambda(\omega_i \mid \omega_j) P(\omega_j \mid x) \tag{3.13}$$

By substituting the posterior probability using Bayes formula the Eq.(3.13) yields to:

$$R(\omega_i \mid x) = \sum_{j=1}^{c} \lambda(\omega_i \mid \omega_j) \frac{p(x \mid \omega_j) P(\omega_j)}{p(x)} \tag{3.14}$$

If the costs are equal $\lambda(\omega_i \mid \omega_j) = \delta_{ij}$

$$\sum_{j=1}^{c} \delta_{ij} \frac{p(x \mid \omega_i) P(\omega_j)}{p(x)} = \frac{p(x \mid \omega_i) P(\omega_i)}{p(x)} = P(\omega_i \mid x) \text{ Bayes MAP} \tag{3.15}$$

Bayesian decision theory offers a clear optimization criterion. Problems that arise are related to the a posteriori probability estimation and likelihood estimation. The likelihood is estimated based on the samples, and this estimation can be parametric, mixt or nonparametric. When the distribution law of data is known the parametric estimation works very well. The mixt method uses a mixture of densities (i.e. Gaussian mixture model or GMM).The problem with this method is related to the number of Gaussians that are needed in order to model the class. The non parametric method estimates the probability density based on the direct observed data or using kernel functions. Partzen introduced for the first time the window method for estimating density functions [187].

3.7 Robust estimation

The robust estimation techniques are very useful in computer vision because here there are even bigger uncertainly regarding the data; especially in feature data (occlusions, different view points).The first step in describing robust estimators is to understand what is meant by robustness. Several measures of robustness are used in the literature one is the breakdown point which represents the minimum fraction of outlying data that can cause an estimate to diverge arbitrarily far from the true estimate. For example, the breakdown point of least squares is 0 because one bad point can be used to move the least squarest arbitrarily far from the true. The theoretical maximum breakdown point is 0.5 because when more than half the data are outliers they can be arranged so that at through them will minimize the estimator objective function.

A second measure of robustness is the influence function[188], [189] which, intuitively, is the change in an estimate caused by insertion of outlying data as a function of the distance of the data from the (uncorrupted) estimate.

Even if is not a measure of robustness, the efficiency of a robust estimator is also significant. This is the ratio of the minimum possible variance in an estimate to the actual variance of a (robust) estimate [190], with the minimum possible variance being determined by a target distribution such as the normal (Gaussian) distribution. Robust estimators having a high breakdown point tend to have low

efficiency, so that the estimates are highly variable and many data points are required to obtain precise estimates.

3.7.1 M-Estimators

The M-estimators are a family of robust techniques which can handle data in the presence of outliers. Their link with kernel density estimators is described in [191]

In computer vision there are two main techniques for robust estimation: M-estimators and least median of squares (LMS). M-estimators are generalizations of MLEs and least squares [188]. It can be defined the M-estimate of a like

$$\hat{a} = \arg\min_{a} \sum_{x_i \in X} \rho(r_{i,a} \mid \sigma_i) \tag{3.15}$$

Where $\rho(u)$ is a robust loss function that grows subquadratically and is monotonically nondecreasing with increasing $|u|$. σ_i^2 is the variance (scale) associated with the scalar value $r_{i,a}$

In order to minimize the Eq (3.15) it is necessary to find a *such* that:

$$\sum_{x_i \in X} \psi(r_{i,a} \mid \sigma_i) \frac{dr_{i,a}}{da} \frac{1}{\sigma_i} = 0 \tag{3.16}$$

Where $\psi(u) = \rho'(u)$ is the influence function. Then a weight function w is introduced; $w(u) \times u = \psi(u)$

$$\sum_{x_i \in X} w(r_{i,a} \mid \sigma_i) \frac{1}{\sigma_i^2} \frac{dr_{i,a}}{da} r_{i,a} = 0 \tag{3.17}$$

This leads to a process known as "iteratively reweighted least squares" (IRLS), which alternates steps of calculating weights $w_i = w(r_{i,a} \mid \sigma_i)$ using the current estimate of *a* and solving Eq. (3.16) to estimate a new *a* with the weights fixed. Initial estimates of *a* may be obtained in a variety of manners, including nonrobust least squares or other robust estimators.

The main differences between M estimators is in the shape of the function $\rho(\cdot)$. The most known robust loss functions are those defined by Beaton- Tukey , Cauchy and Huber. The loss function specifies the amount of outliers that influence the estimate. The shape of the loss function makes a trade-off between efficiency and robustness of the estimator. A narrow function is more robust but less efficient because it might not consider or might use in small percent the inliers data. These functions have a scale factor. The estimation of the scale factor is a major topic of research in robust estimation.

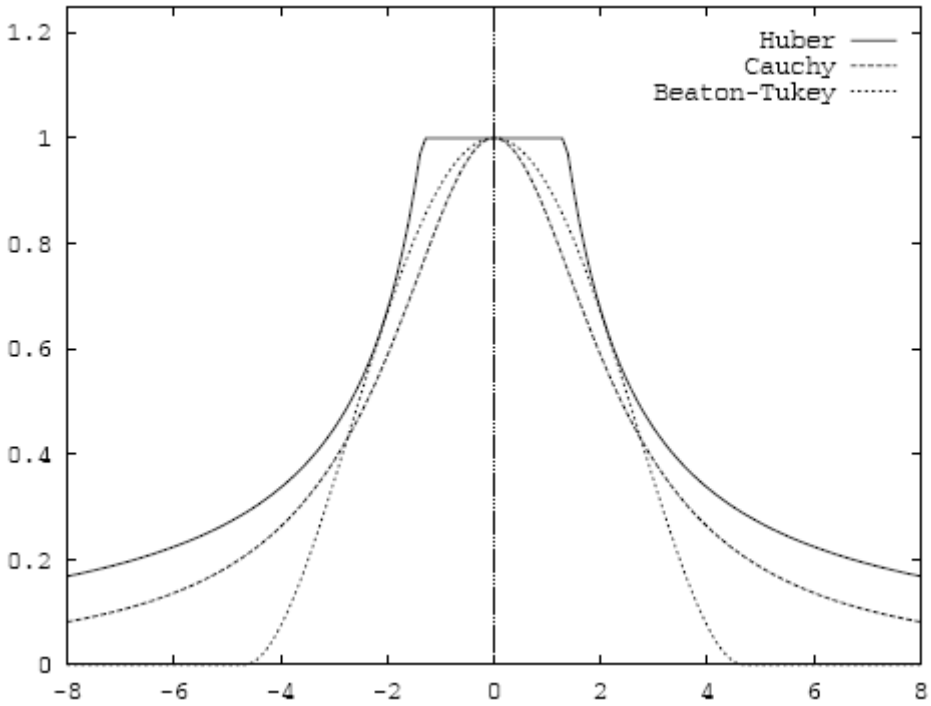


Figure 27 Different shapes for loss functions

4 A COMPOSITIONAL APPROACH TO HAND GESTURE RECOGNITION

4.1 Introduction

In this chapter the compositional technique for hand gesture recognition is presented. The goal of this work is to prove the power of the compositional techniques of hand gesture recognition; the result is presented in a hand gesture recognition system. Compositional techniques have been used with good results in various applications such as: object categorization and data mining. Fei-Fei and Perona [11] used this technique to recognize natural scene categories. R. Fergus [72] learned object categories from Google's image search.

Generally, object recognition approaches consist of four major stages: (1) feature detection, (2) object representation, (3) training, and (4) recognition. This section points out the differences between traditional pattern recognition and compositional approaches with regard to each of these stages.

The first stage – feature detection – uses image regions, interest points, curve fragments, image-filter responses, or a combination of these as image features. Patches, atoms, salient points, interest points, and edges are representative features for sparse representation.

For the second stage – object representation – most approaches partition extracted features into clusters, also known as parts. Our approach is based on *compositions* of parts.

With respect to training, in the third stage, different approaches involve different degrees of supervision in learning object representations. The hand posture is decomposed into relevant compositions which are learned for each hand posture class without supervision; no hand segmentations or localization during training is needed.

Finally, object recognition, in stage four, is typically evaluated only through image classification in terms of whether the learned object class/category is present or absent.

4.2 Overview of the Proposed Approach

In this section an overview of the main steps of our proposed method are presented and their motivation and contribution is pointed out.

Our purpose is to recognize static hand gestures using a compositional technique. Sparse representations are compositional techniques.

Among the most widely used local image features are the interest points, such as: Harris [55], SUSAN (Smallest Univalue Segment Assimilating Nucleus) [180], SIFT (Scale-invariant feature transform) [181], DoG (Difference of Gaussian)[56], [181] etc. While SIFT points are a multiscale generalization of the Harris interest points and proved their usefulness in many modern applications, including object recognition, in this work the Harris feature points were preferred, since the image scale is supposed to be fairly stable. Also the length of the SIFT feature vector (128) is prohibitive for appropriate learning from moderate or small size databases. The region around the Harris interest point is rich in information and

will be described in this work by the colour histogram, the contour orientation histogram, the local direction of interest point and the number of contour points.

Object representation is based on *compositions* of parts: descriptors are grouped according to the perceptual laws of grouping [12] to obtain a set of possible candidate compositions. These groups are a sparse representation of the image based on overlapping subregions.

The detected part descriptors are represented as probability distributions over a codebook which was obtained in the learning phase. A composition is a mixture of the part distributions. From all candidate compositions, relevant compositions must be selected. There are two types of relevant compositions: those compositions that occur frequently in all categories and also those which are specific for a category.

The category posterior of compositions is learned in the training phase, and it is a measure of relevance. The entropy of the category posterior helps us to discriminate between categories. A cost function is obtained by combining the priors of the prototypes and the entropy. The process of recognition is based on bag of compositions method, where a discriminative function is defined.

In Figure 28 can be seen a classical model for statistical pattern recognition according to [192]. In statistical pattern recognition, a pattern is represented by a set of d features, or attributes, viewed as a d-dimensional feature vector. The recognition system is operated in two modes: training and classification. The role of the preprocessing module is to segment the pattern of interest from the background, remove noise, normalize the pattern, and any other operation which will contribute in defining a compact representation of the pattern. The feedback path allows the designer to optimize the preprocessing and feature extraction/selection strategies. In the classification mode, the trained classifier assigns the input pattern to one of the pattern classes under consideration based on the measured features.

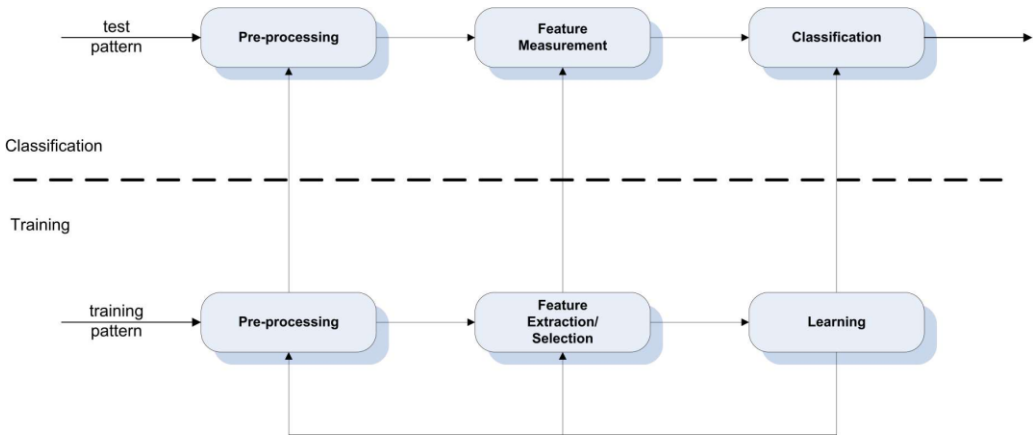


Figure 28 Classical model for statistical pattern recognition

The compositional model for statistical pattern recognition used in this work is shown in Figure 29. The main differences between the classical model for statistical pattern recognition and a compositional model, is that the last one

includes one ore more intermediate layers of abstraction. The features extracted for classical model of pattern recognition have to describe the entire object shape; meanwhile the features selected for compositional techniques are used to generate primitives which combine in order to generate the object shape. There are many ways to generate the object shape, the same primitives combined in different ways conduct to the different shape object, and different primitives combined in a different way can generate the same object shape. The compositional method actually builds a bridge over the semantic gap between low level feature representation and high level of object recognition.

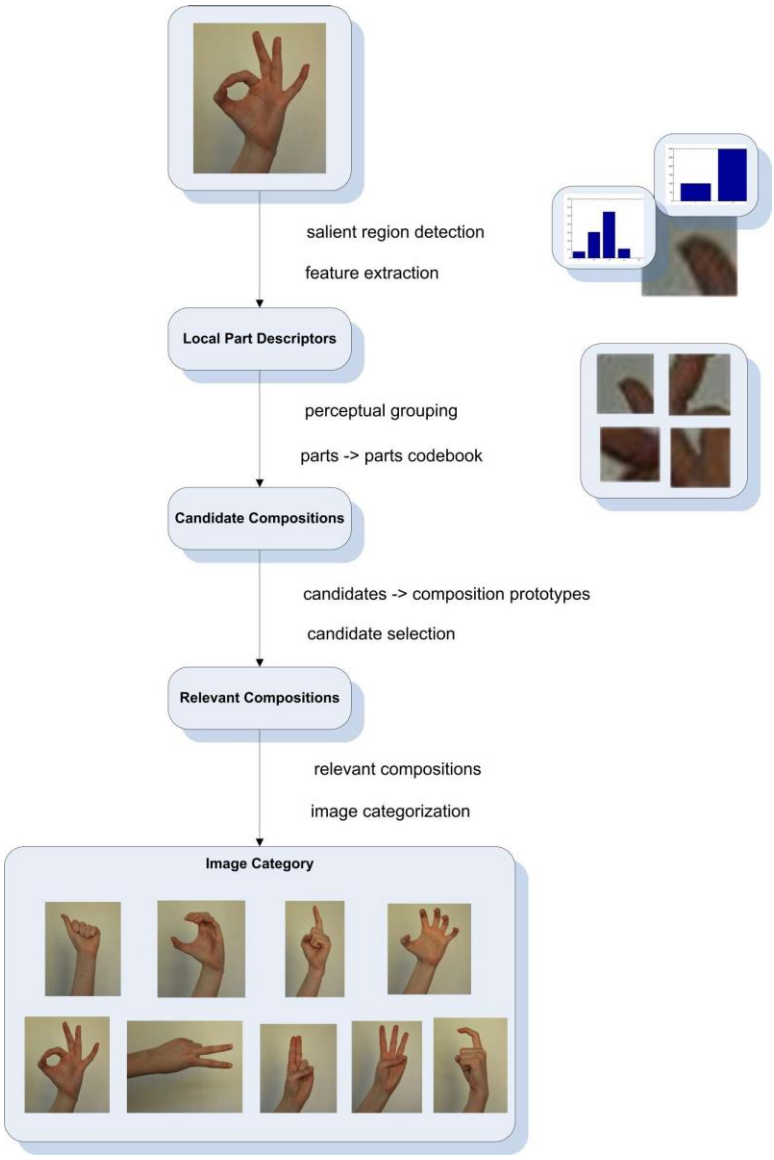


Figure 29 Compositional model for statistical pattern recognition

4.3 Finding good sparse features

4.3.1 Introduction

Even if the proposed method is a general one, for different applications it is still important what features are used for the sparse hand posture representation. The potential benefits of feature selection include, first and foremost, better accuracy of the inference engine and improved scalability. Secondary benefits include better data visualization and understanding, reduced measurement and storage requirements, and reduced training and inference time.

The ideal features are not affected by occlusion and clutter, there are invariant (or covariant), there are also robust, which means that noise, blur, discretization, compression, etc. do not have a big impact on them. From the distinctive point of view individual features can be matched to a large database of objects; from the quantitative point of view many features can be generated for even small objects and offer a precise localization.

The first question according to compositional technique in our case is how hand can be represented in order to be decided which image locations had to be captured and which to dispose of. The main idea is that each hand posture can be described by: the V shapes between the fingers when these are apart, the curve shapes which correspond to the fingertips and the straight lines for the finger length. Each hand pose can be defined as a combination of these shapes. Based on the number of V shapes, curves and lines and based on the relations among them, the hand pose can be recognized. It is important how these shapes are oriented and which their relative position to each other is. The second question is how these relevant image regions can be represented.

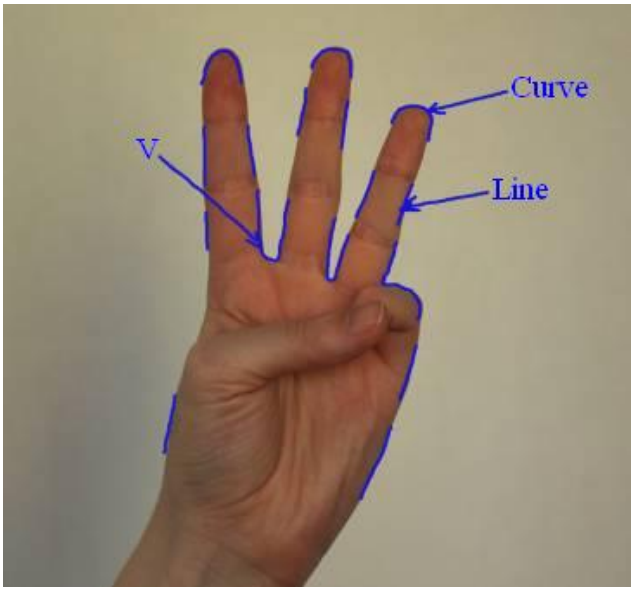


Fig.4.3 V Shape, curve and line

A brief overview of existing detectors is summarized in the following paragraphs.

4.3.2 Interest points and corner detectors

According to Schiele [193] salient points are literally those points on the image which are almost unique. These points maximize the discrimination between the objects.

There is a wide variety of interest points and corners detectors in literature. According to [179] they can be divided into three categories: contour based, intensity based and parametric model based methods.

Contour based methods have existed for a long time[194], [195] and first extract contours, then search for maximal curvature or inflexion points along the contour chains, or do some polygonal approximation and then search for intersection points. Intensity based methods compute [55], [196], [180] a measure that indicates the presence of an interest point directly from the grey values. Parametric model methods fit a parametric intensity model to the signal. Usually these methods provide sub-pixel accuracy, but unfortunately are limited to specific types of interest points.

4.3.2.1 The Harris corner detector

Comparative tests carried out amongst different approaches demonstrated that detectors which combine the smoothed image derivates via the autocorrelation matrix are robust and achieve very good results [179]. Perhaps the most prominent candidate in this category is the edge and corner detector proposed by Harris and Stephens[55]. The Harris corner detector is based on the local auto-correlation function of a signal. The local auto-correlation function measures the local changes of the signal with patches shifted by a small amount in different directions.

4.3.2.2 Difference of Gaussian(DOG)

Difference of Gaussian(DOG) is a popular detector; it was introduced by Lowe in [56], [181]. It defines interest points as the extrema of a convolution of the image with a difference of Gaussian filter. The difference-of-Gaussian representation is obtained by subtracting two successive smoothed images. Thus, all the DoG levels are constructed by combined smoothing and sub-sampling. The local 3D extrema in the pyramid representation determine the localization and the scale of the interest points. For an efficient implementation that works in real-time, the image is repeatedly blurred with Gaussian kernels before taking the difference between successive blurrings. It is scale invariant and has a simple and efficient scheme.

4.3.2.3 Kadir and Brady detector

The *Kadir and* Brady detector [197] searches for scale localized features with high entropy, with the constraint that the scale is isotropic. The algorithm is suitable for finding circular structures. The algorithm generates a space sparsely populated with scalar saliency values. For each pixel location and for each scale value between a minimum and a maxim the local descriptors value is measured within a window; them the PDF from this is estimated and the local entropy is

calculated. The scale which conducts to the peaked entropy is selected. A final clustering of the candidates in scale space does then yield a set of interest points. This method finds regions that are salient over both location and scale.

4.3.3 Local Descriptors.

For the detected regions remains the question: which is the most appropriate descriptor to characterize it? There are many descriptors which emphasize different image properties: pixel intensity, texture, color, edges and others. The simplest descriptor is a vector of image pixels. In praxis some extra processing is needed to reduce the dimensionality and insure invariance to at least limited image transformations. Due to their simplicity this descriptors have been widely used [94], [186], [95]. In [87] once the regions of interest are identified, they are cropped from the image and rescaled to the size of a small typically 11×11 pixel patch. The dimensionality of each patch is 121, and in order to reduce it, the principal component analysis (PCA) is used. In [198] to use the image gradient patch and to apply PCA to reduce the size of the descriptor was proposed. Based on the technique that is used to describe the local image regions according to [199] there are: distribution based descriptors, spatial-frequency techniques, differential descriptors and other techniques.

4.3.3.1 Distribution based descriptors

The scale invariant feature transform (SIFT) proposed by [181] combines a scale invariant region detector and a descriptor base on the gradient distribution in the detected regions. The gradients are then normalized for orientation by rotating the whole region so that its dominant orientation is fixed. A descriptor is a 3D histogram of gradient location and orientation, where location is quantized into a 4x4 location grid and 8 bin orientation histograms are computed for each cell. The resulting descriptor is of dimension 8 x 4 x 4 = 128. Each orientation plane represents the gradient magnitude corresponding to a given orientation. The SIFT features are robust to changes in illumination, noise, and minor changes in viewpoint; they are highly distinctive, relatively easy to extract, allow for correct object identification with low probability of mismatch and are easy to match against a (large) database of local features. To obtain illumination invariance, the descriptor is normalized by the square root of the sum of squared components. The quantization of gradient locations and orientations make the descriptor robust to small geometric distortions and small errors in the region detection. It is used in various applications like object recognition, matching, tracking, and categorization. The high dimensionality and the specificity of these features with respect to individual instances of an object require large codebooks when these features are clustered.

Similar to SIFT features are shape context [58] and geometric histogram descriptor, these also compute a 3D histogram of location and orientation for edge points where all the edge points have equal contribution in the histogram. These descriptors were successfully used for shape recognition of drawings for which edges are reliable features.

Shape context is a 3D histogram of edge point locations and orientations. The object is treated as a point set and the shape of an object is captured by a finite subset of its points sampled from the external/internal contours of the object. The original proposed shape context descriptor was computed only for edge point

locations and not for orientations. In [199] the location is quantized into nine bins of a log-polar coordinate system and orientation is quantized into four bins (horizontal, vertical and two diagonals), the resulting descriptor has a dimension equal to 36.

The geometric blur descriptor [59] is an extension of the shape context descriptor [58]. It blurs the region around an interest point with a spatially varying kernel. Blur should be small near the corresponding points, and larger away from them. Gaussian blur is the right way to simplify image intensity structure and de-noise an image but is not designed with the criterion of making matching points robust under geometric distortions. For a model distortion with affine transformations, the amount of blur varies linearly with the distance from corresponding points. The result is a high dimensional feature vector which describes the image region surrounding an interest point.

Gradient location-orientation histogram (GLOH) is an extension of the SIFT descriptor designed to increase its robustness and distinctiveness. The histogram is computed for 17 location and 16 orientation bins in a log-polar location grid, the higher dimensionality of the descriptor is reduced through principal components analysis to 128.

PCA-SIFT descriptor is a vector of image gradients in x and y direction computed within the support region. The input vector is created by concatenating the horizontal and vertical gradient maps for the 41×41 patch centred at the key-point, having the input vector with 2×39×39=3042 elements. The vector dimension is reduced by performing PCA, the best matching performance was obtained for n = 36, where n represents the dimension space.

4.3.3.2 Spatial-frequency techniques

The Gabor Filters have received considerable attention because they are inspired by the Hubel and Wiesel [60] cells found in the primary visual cortex to some mammals and the Gabor functions seems to be a good model of simple cell receptive fields [200]. Gabor filters are commonly used as an initial representation layer in neuro-physiologically [83] inspired vision systems In addition, these filters have shown to posses optimal localization properties in both spatial and frequency domain and they are well suited for texture segmentation problems. A Gabor filter can be viewed as a sinusoidal plane of particular frequency and orientation, modulated by a Gaussian envelope. The filter is defined by Eq.(4.1):

$$A_{\lambda,\theta,\sigma,\gamma}(x,y) = \exp\left(-\frac{\hat{x}^2+\gamma^2\hat{y}^2}{2\sigma^2}\right)\times\cos\left(\frac{2\pi}{\lambda}\hat{x}\right), \qquad (4.1)$$

$$\hat{x} = x\cos\theta + y\sin\theta$$

$$\hat{y} = -x\sin\theta + y\cos\theta$$

where λ is the wavelength of the filter, θ is its the orientation, σ is the effective filter width controlling the size of the envelope and γ specifies the aspect ratio (the ellipticity of the filter).

4.3.3.3 Differential descriptors

Steerable filters were developed by Freeman and Adelson [61] and describe a class of filters in which a filter of arbitrary orientation is synthesized as a linear

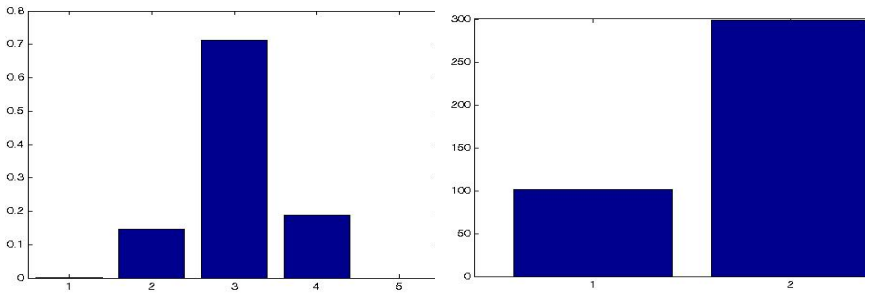
combination of a set of basis functions. The steerable filters steer the derivates in the particular direction given the components of the local derivatives. They are invariant to rotation because they steer the derivatives in the direction of the gradient. All functions that are band limited in angular frequency are steerable, but the most useful are those ones that require a small number of basis filters. A stable estimation of the derivatives is obtained by convolution with Gaussian derivatives.

4.3.4 Conclusions

In this work in order to capture the salient regions Harris corner detector was used on hand contours. The Harris interest point detector is used on hand contours in order to have a low computational cost. It is proved that people can recognize an object from its sketch. Edges are able to capture that information which is enough and useful for our brain-view processor to recognize the object.

Corner features are used because they are local (robust to occlusion) and are relatively stable under certain transformations. It is also claimed that they have high information content. The hand contours were detected by Canny edge detector. One may think that the edge cannot be well detected all the time, and this fact is true, but this approach relays on sparse features, so even if parts of the hand contour are missing, recognition can be done correctly most of the time. In order to describe the region around a Harris interest point, contour localized feature histograms were used. Localized feature histograms were used as a compromise between two opposite goals: perfect localization and maximal invariance aiming at a representation whose invariance properties are transparently adjusted between these two classical extremes and add the specificity lost by invariance through the relations incorporated in compositions.

The orientation histogram of the contour points, number of contours points and the colour histogram were computed. The background should not disturb, so to avoid as much as possible its influence, color histogram with two bins (skin – non skin) are used.



(a) (b)

Figure 30 Example of Orientation Histogram with 4 directions (a) and Colour Histogram with 2 bins (b)

4.4 Detecting sparse features

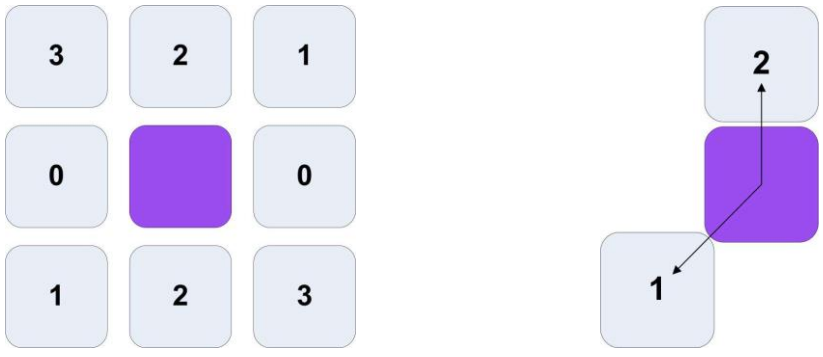
The RGB hand posture image is converted to a gray scale image, and then the Canny edge detector is used in order to extract the hand contours. Salient image locations are detected by using Harris interest point detector on hand contours. Quadratic patches of size 20 × 20 pixels are extracted around each Harris interest point to capture discriminative local information. The patch size is choose to capture the fingertip. For each extracted patch its correspondent in the RGB image is searched and a two bin color histogram (skin-non skin) is extracted.

Numerous colourspaces for skin modeling have been proposed. The most popular color spaces are: RGB, HIS, HSV, HSL, TLS, YCrCb [201]. RGB is a colourspace originated from CRT display applications, when it was convenient to describe color as a combination of three colored rays (red, green and blue). It is one of the most widely used colourspaces for processing and storing of digital image data.

Hue-saturation based colourspaces were introduced when there was a need for the user to specify color properties numerically. They describe color with intuitive values, based on the artist's idea of tint, saturation and tone. Several interesting properties of Hue were noted in [202]: it is invariant to highlights at white light sources, and also, for matte surfaces, to ambient light and surface orientation relative to the light source and hue is also less sensitive to different skin colour. RGB, HS and Hue colour spaces have shown the best results for colour skin segmentation.

In this work the goal of the 2 bin color histogram is to detect different types of regions around the interest point assuming that the background is extracted. The background extraction can be done using one of the proposed methods in literature [203].

The contour orientations histogram with four bins is also extracted. Using Canny edge detector the obtained contours are thin and each contour point contribute to the histogram with one, two (as it can be seen in Figure 31 (b)) or more local directions.



(a) (b)

Figure 31 The 4 orientations considered (a) Contour points contribution to histogram (b)

We define:

$$N(i) = \begin{cases} 1, EDGE = True \\ 0, Else \end{cases} \tag{4.2}$$

The orientation of a contour point is defined:

$$O_{x,y}(i) = \begin{cases} i, N(i) = 1 \\ 0, Else \end{cases} \tag{4.3}$$

The histogram is computed according to Eq.(4.4)

$$h(i) = \sum_{x,y \in region} O(i) \tag{4.4}$$

Then the relative direction of the interest point is computed similar with contour orientation histogram. For the same patch the number of contour points is also extracted.

Figure 32 Examples of Harris interest points detected on hands contours

The resulting eight parameters extracted from a patch are used to form a feature vector, $\mathbf{e}_i$.

Figure 33 Fingertip region

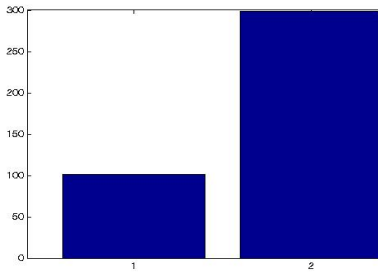


Figure 34 Color Histogram with 2 bins for the fingertip region

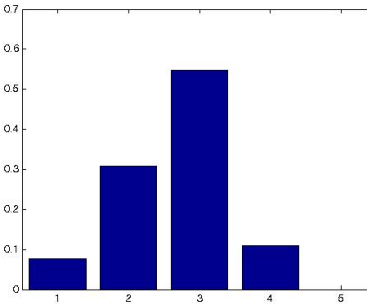


Figure 35 Orientation Histogram with 4 directions for the fingertip region

Figure 36 V region

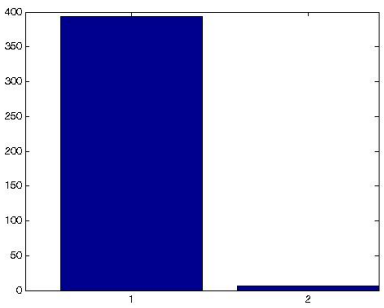


Figure 37 Color Histogram with 2 bins for the V region

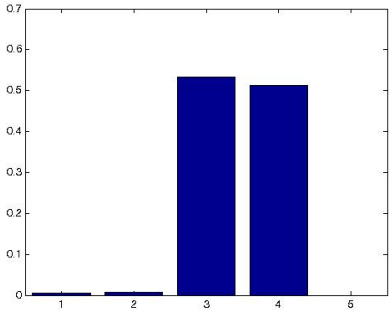


Figure 38 Orientation Histogram with 4 directions for the V region

Figure 39 Line region

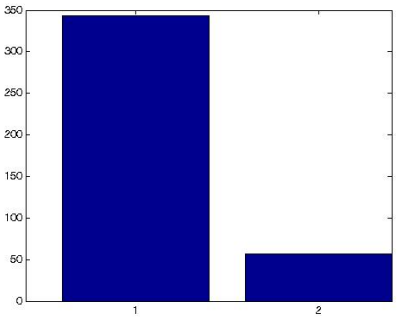


Figure 40 Color Histogram with 2 bins for the line region

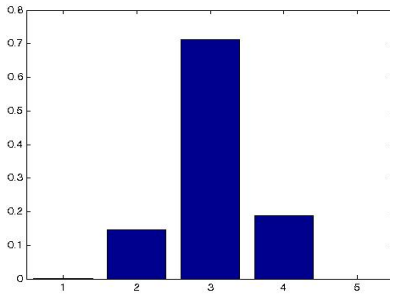


Figure 41 Orientation Histogram with 4 directions for the line region

It is important to remark the small dimension of the feature vector, which is eight. An example of interest points detected with Harris interest point detector on the hand contour, extracted with Canny edge detector can be seen in Figure 32.

The Harris interest point detector is used on hand contours in order to have a low computational cost and as it was previously stated edges are abele to capture that information which is enough and useful for our brain-view processor to recognize the object.

4.5 Hand posture representation

Hand posture representation is based on *compositions* of parts. In the proposed approach, a part is an image patch around a Harris interest point, described by a feature vector $\mathbf{e}_i$. All feature vectors $\mathbf{e}_i$ from all images from a training set are clustered using k-means, in order to generate a codebook with *relevant features for all hand posture classes*. The codebook is subsequently used in order to assess the similarity of extracted image features to learned classes of relevant features. Notice that feature classes obtained by clustering are not related in any way to hand posture classes. Instead, feature classes in a compositional approach are used to generate an alternative representation of image parts, as explained later on.

4.5.1 Generating a codebook of relevant features

Data clustering is a generic label for a variety of procedures designed to find natural groupings, or clusters, in multidimensional data, based on measured or perceived similarities among the patterns. The problem is difficult because data can reveal clusters with different shapes and sizes. Accordingly different clustering methods may be more appropriate to discover the best grouping corresponding to the purpose of data analysis.

The results of a data clustering method mainly depend on two aspects: the distance measure and the grouping strategy. Arguably, the most important step in any clustering method is the selection of the distance measure. This measure quantifies the similarity or dissimilarity between data points or data points and cluster centers. This measure will also influence the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another.

Cluster analysis is a very important and useful technique, owing to the speed, reliability, and consistency with which a clustering algorithm can organize large amounts of data. The clustering algorithms are used in various applications like: data mining [204], information retrieval [205], image segmentation [206], signal compression and coding [207], and machine learning [208]. As a consequence, hundreds of clustering algorithms have been proposed in the literature like: K-means, Fuzzy K-means, Minimum Spanning tree, Mutual Neighborhood, Single-Ling, Complete-Link, Mixture Decompozition, and new clustering algorithms continue to appear. According to [192], there are two popular types of clustering techniques: agglomerative hierarchical clustering and iterative square-error partitional clustering. Algorithms for hierarchical clustering are either agglomerative, or divisive. Hierarchical techniques organize data in a nested sequence of groups which can be displayed in the form of a dendrogram or a tree.

The agglomerative algorithms use a bottom up approach, each observation starts in its own cluster, and pairs of clusters are merged as one move up the hierarchy. The root of the tree consists of a single cluster containing all observations, and the leaves correspond to individual observations. The divisive ones are using a top down approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

It is hard to say which clustering algorithm is the best, for a given dataset. The best way to decide is to try several clustering algorithms. All the issues related to data collection, data representation, normalization, and cluster validity are as important as the choice of clustering strategy.

The simple K-means partitional clustering algorithm is computationally efficient and gives unexpected good results, if the clusters are compact, hyperspherical in shape and well-separated in the feature space. The algorithm has 3 steps the first step selects an initial partition with K clusters, the second generates a new partition by assigning each pattern to its closest cluster center; the third step computes new cluster centers as the centroids of the clusters; the step two and three repeats until an optimum value of the criterion function is found.

The k-means clustering algorithm can be identified to be a particular case of the EM (expectation maximization) algorithm. It is similar to the expectation-maximization algorithm for mixtures of Gaussians in that they both attempt to find the centers of natural clusters in the data. The expectation-maximization algorithm maintains probabilistic assignments to clusters, instead of deterministic assignments.

The meanshift clustering technique requires no prior knowledge of the number of clusters, and does not constrain the shape of the clusters. The mean shift vector always points toward the direction of the maximum increase of the density. The mean shift procedure, obtained by successive computation of the meanshift vector and translation of the window with meanshift vector define in Eq.(4.5) is guaranteed to converge to a point where the gradient of density function is zero.

$$\mathbf{m}_{\mathrm{h},\mathrm{G}}(\mathrm{x}) = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathbf{x}_{\mathrm{i}} g\left(\left\| \frac{\mathbf{x}-\mathbf{x}_{\mathrm{i}}}{\mathrm{h}} \right\|^2\right)}{\sum_{\mathrm{i}=1}^{\mathrm{n}} g\left(\left\| \frac{\mathbf{x}-\mathbf{x}_{\mathrm{i}}}{\mathrm{h}} \right\|^2\right)} - \mathbf{x} \tag{4.5}$$

In Eq.(4.5) h is the bandwidth parameter, $\mathbf{x}_{\mathrm{i}}$ represent the data, $\mathbf{x}$ is the window center, and g is the kernel profile.

Both K-means and meanshift clustering techniques have been tried on the dataset. The best results, as it can be seen in next chapter, have been achieved by performing K-means clustering. This result confirms the optimality of dominant approaches described in literature [10],[186].

The number of clusters *k,* for this application, was set to five. The main reason why the number of clusters is five is related to the types of patches detected in an image. A patch from the image may be characterized by:

- more skin region and less background region,
- more background region and less skin region,
- the skin and background region may have the same percentage in the same patch,
- the patch may have only background respectively

- the patch may have only skin regions.

To make the representation more robust and to make it less susceptible to local minima in the expectation-maximization (EM) iterations of k-means, each feature point is described by a Gibbs distribution over the codebook like in [209] instead of being simply labeled with the class label of its nearest prototype. In order to alleviate the local minima problem k-means can be run with several initializations and the best solution is selected.

Gibbs distributions are characterized by their energy functions and these are more convenient and intuitive for modeling than working directly with probabilities. In addition, the Gibbs distribution is the unique measure that maximizes the entropy for a given expected energy [210]. The probability measures from Eq 4.4 are always positive and hence random fields.

According to the Gibbs distribution law, the feature assignment random variable, F_i, is given by Eq. (4.6).

$$P(F_i = v \mid \mathbf{e}_i) = Z(\mathbf{e}_i)^{-1} \exp(-d_{v,\sigma}(\mathbf{e}_i)) \tag{4.6}$$

$$Z(\mathbf{e}_i) = \sum_v \exp(-d_{v,\sigma}(\mathbf{e}_i)) \tag{4.7}$$

$$d_{v,\sigma}(\mathbf{e}_i) = \left\| \mathbf{e}_i - \mathbf{a}_v \right\|^2 \tag{4.8}$$

where F_i is a feature assignment random variable, $P(F_i = v \mid \mathbf{e}_i)$ is the probability of feature vector $\mathbf{e}_i$ to belong to the class v defined by the prototype vector $\mathbf{a}_v$, $d_v(\mathbf{e}_i)$ is the Euclidian distance of a measured feature $\mathbf{e}_i$ to a centroid $\mathbf{a}_v$ of class v and σ is a normalization factor. Eq.(4.6) is evaluated for all centroids $\mathbf{a}_v$ and the results for a feature point described by $\mathbf{e}_i$ are grouped in a *part distribution vector*:

$$d_i = \left(P(F_i = 1 \mid \mathbf{e}_i), \ldots P(F_i = k \mid \mathbf{e}_i)\right)^T \tag{4.9}$$

4.5.2 Generating compositions of image parts

In order to form a higher level of abstraction, image parts are grouped into compositions. In order to decide which parts should be grouped to form the candidate compositions the principles of perceptual organization are used. To this end, all detected local parts from an image, represented by their part distribution vector, are grouped with their neighbours that are not farther away than *N* pixels. This grouping principle follows the principle of perceptual organization from Gestalt laws, more precisely the grouping *principle of proximity [12]*. In this work the number of pixels *N* is 25. This number depends on the types of objects and compositions that one wants to form, the number of interest points detected in an image, the number of objects present in an image and also by the image resolution. In [10] this number is between 60-100 pixels.

Gestalt psychology is a theory which refers to the visual perception developed by German psychologists in the 1920s. This theory attempts to describe how people tend to organize visual elements into groups or unified wholes; this theory proposes that the operational principle of the brain is holistic, parallel, and analog, with self-organizing tendencies; or, that the whole is different from the sum of its parts. The form-forming capability of our senses is the effect this theory refers to. The Gestalt psychology was applied to visual recognition of figures and whole forms instead of just a collection of simple lines and curves.

Candidate compositions are represented as mixtures of the part (feature point) distributions as defined in Eq. (4.6). If $\Gamma_j = \{\mathbf{e}_1, ...\mathbf{e}_{mj}\}$ denotes the grouping of parts represented by $\mathbf{e}_1, ..., \mathbf{e}_{mj}$, and $\mathbf{d}_1, ..., \mathbf{d}_{mj}$, (where m is the number of vectors which generate the candidate composition), compositions are then represented by the vector valued random variable G_j which is a bag of parts with the particular values given by:

$$\mathbf{g}_j \alpha \frac{1}{m} \sum_{i=1}^{m} \mathbf{d}_i = \frac{1}{m} \sum_{i=1}^{m} \left(P(F_i = 1 \mid \mathbf{e}_1), ... P(F_i = k \mid \mathbf{e}_i) \right)^T \tag{4.10}$$

where the number of constituents, $m_j = |\Gamma_j|$, is not predefined and can be different for each composition. It depends on how many parts the grouping algorithm can combine into composition in a certain region of an image. Note that the representation of a composition depends on the type of constituent parts and not on the number of parts. A composition is represented by the vector $\mathbf{g}_j$, which can be thought of as the average distribution of its parts over the codebook containing relevant parts for recognition. This model is also robust with respect to variations in the individual parts.

4.5.2.1 Learning relevant compositions

On the set of all compositions that can be formed, a selection of *relevant compositions* must be performed in order to have the discriminative ones and to discard the clutter. The relevant compositions must reflect a trade-off between *generality* and *singularity*. The goal is to learn a small number of compositions so that estimating category statistics on the training data becomes feasible. There are compositions which are present in many classes and there are compositions that help to discriminate sets of classes from another, not necessarily one class from all the other.

First, compositions which are specific for a large majority of hand posture classes are learned. These compositions should be shared among many classes. In order to do this, in the learning phase, all composition candidates found in all the training images, represented by average distribution vector of parts, $\mathbf{g}_j$, are clustered using once more k-means clustering. Let $\pi_i \in \Pi$ be the composition prototypes found by clustering. Then the prior assignment probabilities of candidate compositions to clusters $P(\pi_i)$, are computed using the Gibbs distribution:

$$P(\pi_i = \Pi \mid \mathbf{g}_j) = Z(\mathbf{g}_j)^{-1} \exp(-d_{\Pi,\sigma}(\mathbf{g}_j)) \tag{4.11}$$

$$Z(\mathbf{g}_j) = \sum_{\Pi} \exp(-d_{\Pi,\sigma}(\mathbf{g}_j)) \tag{4.12}$$

In the second stage, relevant composition prototypes for specific classes are selected. Those prototypes help to distinguish between classes. To this end, the category posteriors of compositions must be estimated. In order to estimate the category posteriors of compositions a Bayesian approach was used:

$$P(c \mid \Gamma_j) = \frac{P(\Gamma_j \mid c)P(c)}{P(\Gamma_j)} = \frac{P(\Gamma_j \mid c)P(c)}{\sum_c P(\Gamma_j \mid c)P(c)}$$
$$P(c \mid \Gamma_j) \approx \frac{P(\Gamma_j \mid c)}{\sum_c P(\Gamma_j \mid c)} \tag{4.13}$$

where $c \in \wp$, $\wp$ is the set of all category hand postures. We assume that P(c) are equal, all classes are used with the same probability. The category posterior is used to calculate the relevance of a composition for discriminating hand postures. In order to find a relevance measure the category posteriors of compositions are learned from the training data. The relevance of a composition for discriminating hand postures is then estimated by the entropy of its category posterior:

$$H(P_{\Gamma_j}) = -\sum_{c \in \wp} P(c \mid \Gamma_j) \log P(c \mid \Gamma_j) \tag{4.14}$$

The entropy is used as a measure of discriminative relevance; since entropy measures how uniformly a random variable is distributed the entropy should be minimized.

In order to measure the total relevance of a compositional prototype, a cost function is defined. The cost function combines the prior assignment probabilities of clusters and the entropy, so it combines the reusability criterion with the criterion that measures the ability of compositions to discriminate hand postures from one another. The resulting cost function defined guides the selection of relevant compositions.

There are two cost function proposed in literature by the same authors. In [10] Ommer and Buhman proposed the following cost function:

$$S(\boldsymbol{\pi}_i) \alpha - P(\boldsymbol{\pi}_i) + \lambda H(P_{\boldsymbol{\pi}_i}) \tag{4.15}$$

In [126] Ommer proposed the cost function described by Eq.(4.16).

$$S(\boldsymbol{\pi}_i) \alpha - \log P(\boldsymbol{\pi}_i) + \lambda H(P_{\boldsymbol{\pi}_i}) \tag{4.16}$$

Both constituents of the cost function should be normalized to the same dynamic range, giving rise to an additional additive constant that can be discarded and to the parameter λ.

4.5.2.2 Robust approach to parameter selection in relevant prototype set generation.

Parameter λ defines the balance between the two conflicting demands: generality and specificity. Its value proved to be very important in practice. Parameter λ reflects the way the generality and specificity combines in order to select the relevant prototypes which determinates farther the relevant composition used to describe an image. In [126] it is proposed a method to compute the value of the parameter λ. The estimation of parameter λ proposed by Ommer is not a robust one because it uses the maximum and minimum values which are sensitive to outliers, as one can see in Eq.(4.17).

$$\lambda = \frac{\max_i \log P(\pi_i) - \min_i \log P(\pi_i)}{\max_i H(P_{\pi_i}) - \min_i H(P_{\pi_i})} \tag{4.17}$$

In this approach the parameter is estimated using the inter-quartile range (IQR) which is equal to the difference between the third and first quartiles. A **quartile** is any of the three values which divide the sorted data set into four equal parts, so that each part represents one fourth of the sampled population. The inter-quartile range gives a measure of the spread represented by half of the entire sample and has the advantage of excluding extreme values, so the inter-quartile range is a robust estimator. The proposed robust method for estimating parameter λ is presented in Eq. (4.18).

$$\lambda = \frac{IQR(P(\pi_i))}{IQR(H(P_{\pi_i}))} \tag{4.18}$$

From the set of all compositional prototypes a set of relevant composition prototypes is established through minimization of Eq.(4.15). For all composition prototypes $\boldsymbol{\pi}_i$ the cost function is computed and a set of *r* relevant composition prototypes is selected. The distance between all composition and all relevant composition prototypes and irrelevant compositional prototypes is computed. The image is represented by those candidate compositions which are closer to the relevant prototypes than any irrelevant ones.

The result obtained using the proposed equation for λ computation is better than one obtained using Ommer equation. The results are presented in the next chapter.

4.6 Training step

The training procedure is carried out according to the diagram shown in Figure 43. For all training images the features vectors $\boldsymbol{e}_i$ are extracted and K-means is performed in order to generate the feature codebook, which is the first product of the training step. Based on feature vectors and the feature codebook, the candidate compositions are extracted and modeled with their distribution vectors over the feature codebook.

Candidate compositions from all test images are clustered using one more time k-means, and the resulted composition prototypes are used to form the composition codebook. Based on the cost function defined in Eq. (4.15) *relevant* compositions prototypes are learned in the next stage. A set of *r* relevant composition prototypes is established. This set is obtained by selecting the prototypes $\boldsymbol{\pi}_i$ with minimal cost $S(\boldsymbol{\pi}_i)$. Only those relevant compositions which are not farther away from the relevant composition prototypes than the irrelevant ones are retained.

Each image from the training set is described by those candidate compositions which are closer to the relevant prototypes than any irrelevant ones (these are the relevant compositions) and also by the relative rescaled position coordinates of the relevant compositions.

The hand position may vary from one image to another, so in order to get invariance to translation the relative coordinates are used. The relative position of the compositions is estimated using the median, not the mean because the median is more robust. These relative positions are rescaled using parameter α. In Figure 42 an example of relevant composition detected for class f are shown.

Figure 42 Example of relevant compositions

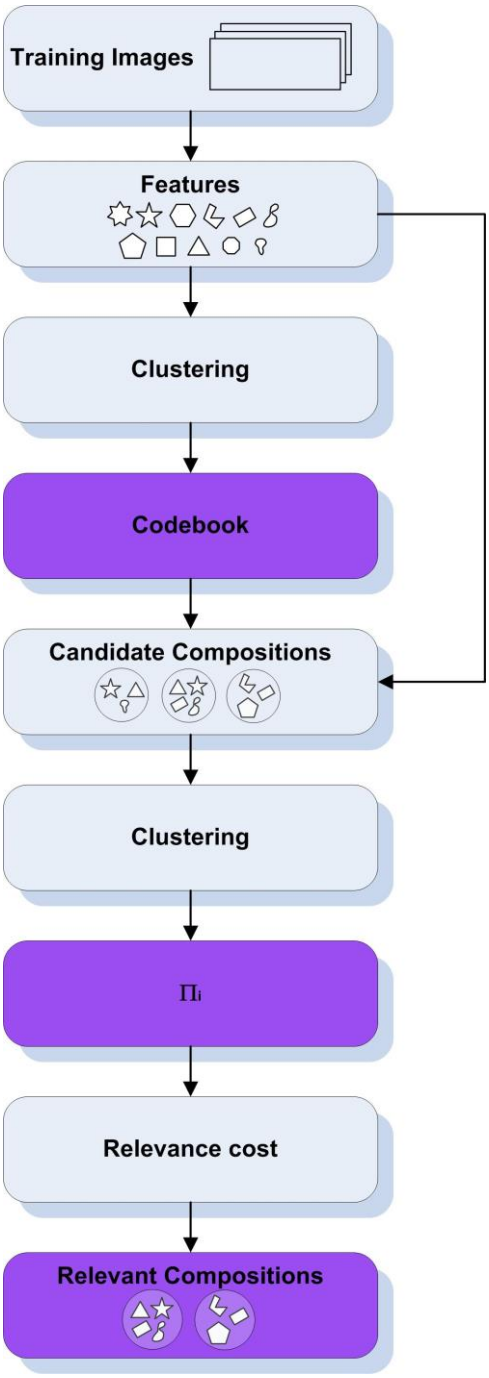


Figure 43 Training diagram

4.7 Hand posture Recognition

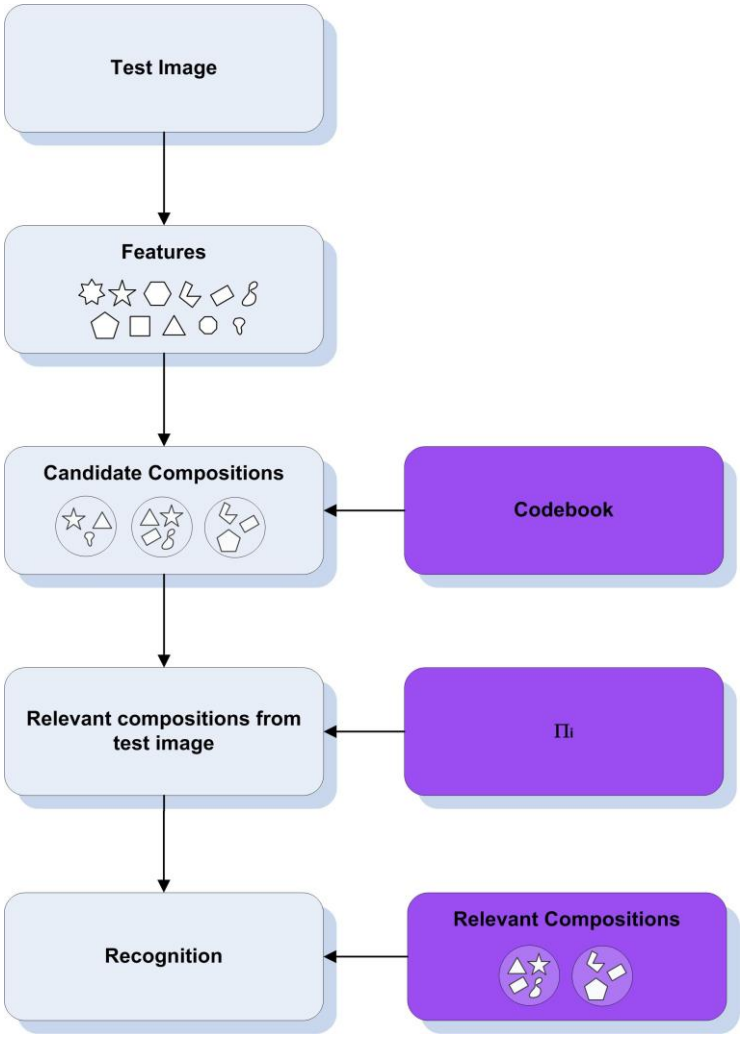


Figure 44 Recognition step

The recognition part is done based on the bag of compositions method. For the new image, a set of composition vectors $\mathbf{h}_i$ is computed. These vectors consist of $\mathbf{g}_j$ distributions and relative, rescaled position coordinates of the relevant compositions. In order to get invariance to translation the relative rescaled coordinates x_i, y_i are used. Hand position is estimated using the median, not the mean because the median is less influenced by the maximum and minimum values from the set of coordinates and is more robust. Evaluation of the data set using median is good if half of the data is correct. For this application more than half of

the data is correct because most of the compositions are generated from interest points located on hand and less from interest points found on background.

The relative position is rescaled using the parameter α. The evaluation of the parameter α is a problem of feature extraction and depends on the data characteristics. Its value influences the space shape.

$$\mathbf{h}_i = \begin{bmatrix} x_i \\ y_i \\ \mathbf{g}_j \end{bmatrix} = \begin{bmatrix} \alpha x_r \\ \alpha y_r \\ \left(P(F_i = 1 \mid \mathbf{e}_1), \ldots P(F_i = k \mid \mathbf{e}_i)\right)^T \end{bmatrix} \tag{4.19}$$

Where $x_i = \alpha(x - x_{median}) = \alpha x_r$

$y_i = \alpha(y - y_{median}) = \alpha x_r$

The range for x_r, y_r is larger than the range of probabilities. Both compositions and their position should have similar importance because the hand posture is recognized based on types of compositions and their relative position one to another. The value of parameter α is learned based on the experimental data.

4.7.1 Hand posture classification

The classification of a new image which is described by vectors $\mathbf{h}_i$ is not straight forward. The number of compositions that describe the testing image differs from the number of compositions which describes the images from the bag (each image from the bag might have different numbers of compositions). All components which describe an image can be seen as a vector; because the length of the vectors is not equal for all images it is not possible to use traditional classifications methods, for example neural-networks [103].

The proposed classification method is inspired by point matching used in image registration, where two sets of points need to be registered and correspondence of points need to be formed. The two sets of points usually suppose different numbers of points. The minimum distance from a fixed point $a_i \in A$ found in set 1, to points $b_n \in B$ from set 2 (according to Figure 45) is shown in Eq.(4.20)

$$\min_{\forall n} d(a_i, b_n) = d(a_i, b_l) \tag{4.20}$$

The minimum distance from point $b_l \in B$ to points from set 1 according to Figure 45 is:

$$\min_{\forall n} d(b_l, a_n) = d(b_l, a_k) \tag{4.21}$$

In Figure 45 it can be seen that $d_1 \neq d_2$, where $d_1 = d(a_i, b_l)$ and $d_2 = d(b_l, a_k)$.

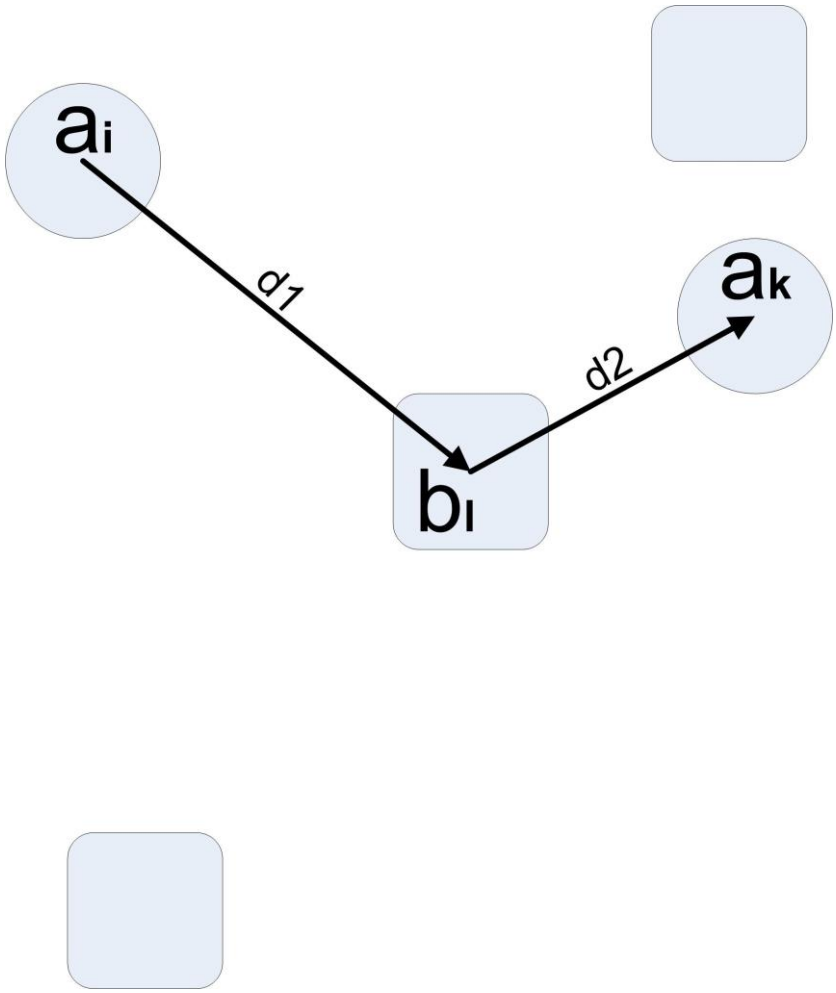


Figure 45 Minimum distance between points.

For each new image only the minimum distance from the training images compositions to test image composition $\mathbf{h}_i$ is computed $\min_{c_i} \left\| h_v^{k,q_v} - h_i^{c_i} \right\|$, then all these distances are sum and normalized according to Eq.(4.22)

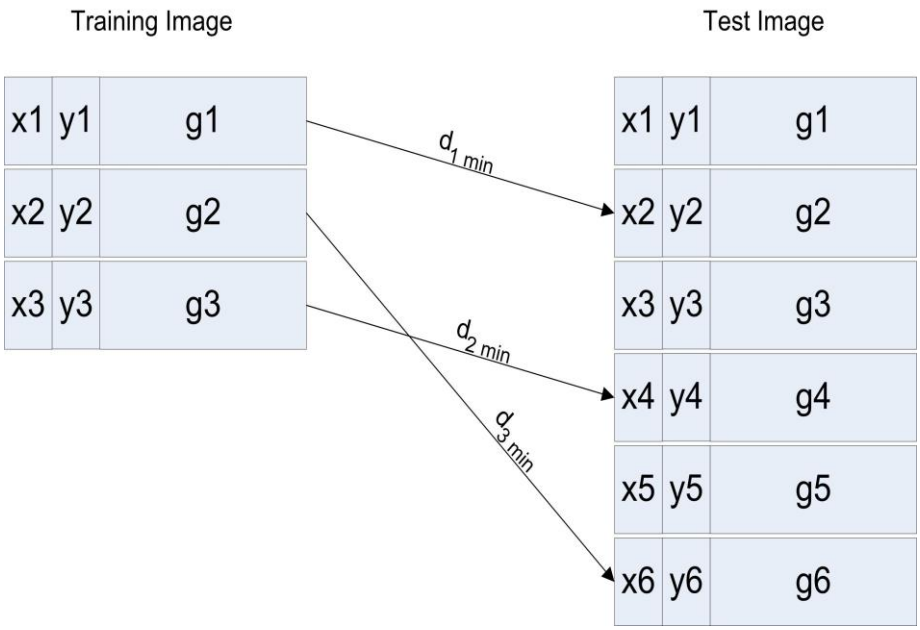


Figure 46. Minimum distance between compositions found in training image and those found in the test image.

v is the number of pictures per class, k is the class, q_v is the number of compositions from a class, i is the current image and c_i is the number of composition for the test image.

$$d(c, v_k) = \frac{1}{\#q_v} \sum_{q_v} \min_{c_i} \left\| h_v^{k,q_v} - h_i^{c_i} \right\| \tag{4.22}$$

$$d(c,k) = \underset{v_k}{\operatorname{argmin}}\, d(c, v_k) \tag{4.23}$$

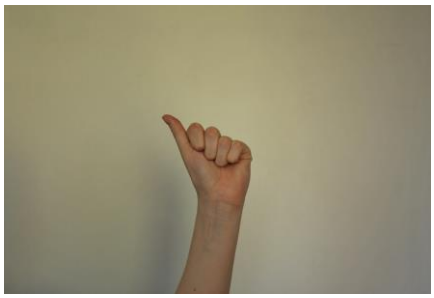
The reason why the distance from the test image compositions to training images is not computed is related to the fact that the testing image might have some compositions which are not specific for that class; it might have compositions as a result of some interest points detected on background. This is less likely to happen for training images.

These distances are computed for all images.

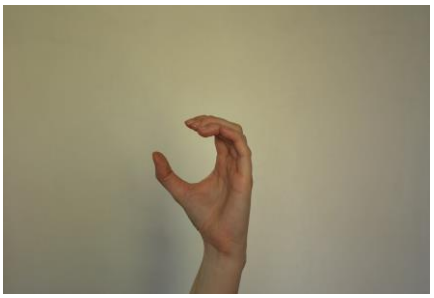
The discriminant function used in the experiments from this work is defined as:

$$k_{opt} = \underset{k}{\operatorname{argmin}}\, d(c,k) \tag{4.24}$$

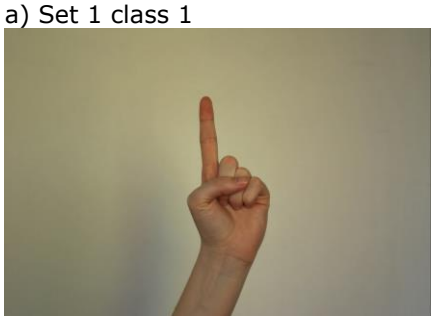
In order to prove the power of the compositional approach in hand posture recognition, two sets of hand gesture were used. The first one consists of nine classes of hand postures and the second one is represented by six classes as it can be seen in Figure 47 and Figure 48



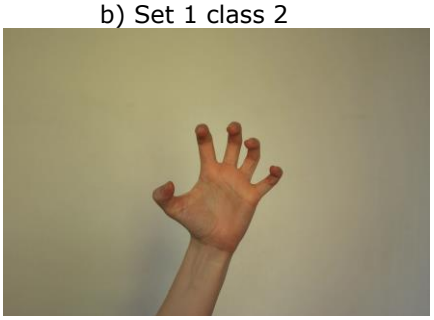
a) Set 1 class 1



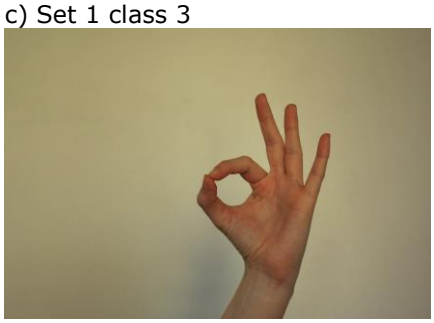
b) Set 1 class 2



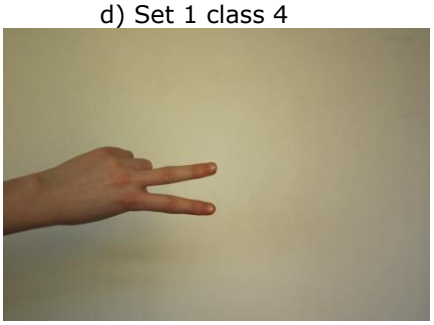
c) Set 1 class 3



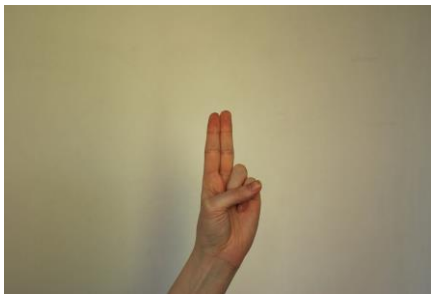
d) Set 1 class 4



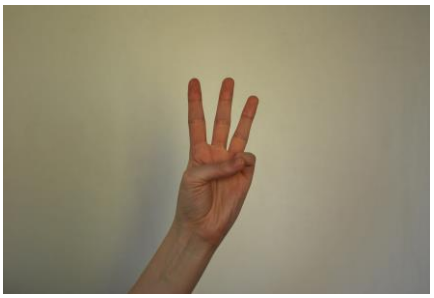
e) Set 1 class 5



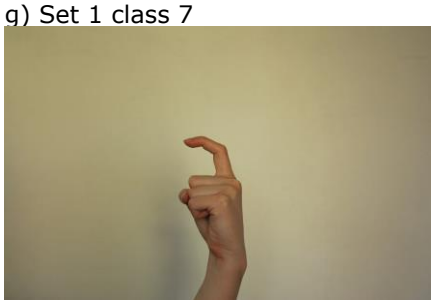
f) Set 1 class 6



g) Set 1 class 7

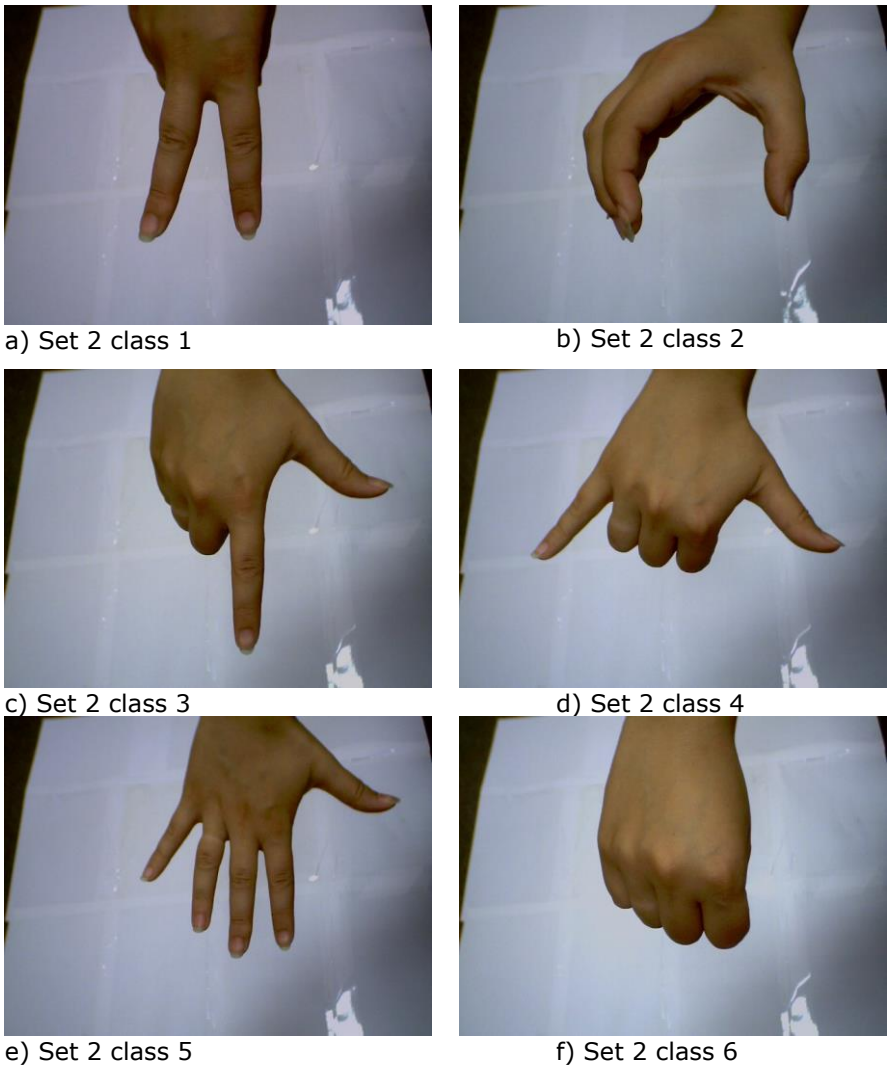


h) Set 1 class 8



i) Set 1 class 9

Figure 47 Training set 1 with 9 classes



a) Set 2 class 1 b) Set 2 class 2

c) Set 2 class 3 d) Set 2 class 4

e) Set 2 class 5 f) Set 2 class 6

Figure 48 Training set 2 with 6 classes

4.8 Conclusions

In this chapter a compositional approach for hand posture recognition was described. The goal of this work was to prove the power of compositional techniques in hand gesture recognition. The compositional techniques have been used with good results in applications like: object categorization and data mining; however these techniques have not been used in classification. The main advantage of the compositional techniques is their generality; these techniques are more independent

of application. Compositional techniques are well suited to incorporate principles from the Gestalt theory of visual perception, therefore they have an important and mostly unexplored potential for further development. Gestalt theory tends to emulate better the way our brain-view processor works. Nowadays research in human vision makes us understand more about the process that people use to recognize an object and this helps the Computer Vision Community develop more similar techniques to the human vision. This work is an attempt to extend the types of problems solved based on the new, compositional approach. While using the general framework of some reference compositional techniques [209], this work designed the processing modules by considering the specifics of the hand gesture recognition problem, where needed.

The main contribution of this work is the compositional approach used to hand posture recognition. According to compositional technique first is decided how hand can be represented in order to know which image locations have to be captured and which to dispose of. The idea is that each hand posture can be described by: the V shapes between the fingers when these are apart, the curve shapes which correspond to the fingertips and the straight lines from the finger length. Each hand pose can be defined as a combination of these shapes. Based on the number of V shapes, curves and lines and based on the relations among them the hand pose can be recognized.

One of the contributions of this work is to carefully select the basic features (contours, interest points, patches, colour histograms, orientation histograms). These basic features generate the primitive features (the V shape, the curves and the lines). The primitive features are like Lego components, they are not extremely variated but by combining them it is possible to generate a lot of object shapes. A contribution of this work also consists in the selection of primitive features.

The object representation is based on *compositions* of parts: descriptors are grouped according to Gestalt law of proximity to obtain a set of possible candidate compositions. **In order to generate the desired primitive features it was important to choose the right distance between the parts which are about to be grouped.** *Candidate* compositions from all test images are clustered and the resulted composition prototypes are used to form the composition codebook. Based on the cost function the relevant compositions prototypes are learned in the next stage. **The optimization of parameter λ, its robust estimation in order to select the relevant compositions prototypes represents a major asset of this work**.

Based on relevant compositions prototypes the relevant compositions are selected. Relevant compositions and their rescaled position is used to describe the image. **Both relevant compositions and their position should have similar importance because the hand posture is recognized based on types of compositions and their relative position one to another. In order to have this, the parameter α is introduced and its value is learned based on the experimental data.**

The discriminant function for classification inspired by the point matching used in image registration, represents also a contribution of this work.

5 EXPERIMENTS

In this chapter the results of the experiments which evaluate the proposed compositional technique for hand gesture recognition are presented.

5.1 Experimental settings

In order to evaluate the compositional approach to hand gesture recognition two sets of hand postures are used. The first one consists of nine classes of hand postures and the second one is represented by six classes of hand postures. The second set of hand postures has six classes other then those from set 1. The first set of hand postures represents nine hand postures from ASL (American Sign Language) and the second one represents six classes of hand postures that can be perform pretty easily in front of a webcam while a person is sited on a chair. For the first set there were taken pictures with Nikon D60 and for the second one a Canyon web cam was used. In figures 49 to 57 different hand postures from different classes can be seen.

For the first set of hand postures 30 training images per class are used and for the second set 60 training images per class are used. The first set of training images has as background a white wall. The second set of training images has as background a white paper. Both training set pictures are taking in natural conditions, no artificial light was added. For the first set of images there is a single subject, the same is for the second set, but the subjects are different.

5.2 Experiments results for set 1

The first set of hand postures represents nine hand postures from ASL (American Sign Language), more exactly letters: a, c, d, e, f, p, u, w, x. The pictures from set 1 are taken in two different days; this can be notice from illumination changes, which is not the same for all pictures.

For set 1 the images resolution is 255 px x 171px.

In order to extract the hand contours Canny edge detector from Image Processing Toolbox, Matlab is used. The sensitivity thresholds for the Canny method for set 1 is defined; the high threshold *thresh* is 0.5 and 0.4**thresh* is used for the low threshold.

The number of clusters *k*, is five in all experiments, and the number of composition prototypes $\boldsymbol{\pi}_i \in \Pi$ varies in experiments.

The two bin colour histograms are extracted only for the red component of the RGB image. Parameter σ from Eq.(4.6) is equal to 1 and parameter σ from Eq.(4.11) equal to 0.05.

The number of composition prototypes $\boldsymbol{\pi}_i \in \Pi$ is 20 and the number of relevant composition prototypes *r*, which conduct to the best result is equal to 19. The number of relevant prototypes is 19 because almost all compositions resulted from interest points detected on hand and just a few are the result of some points detected on background.

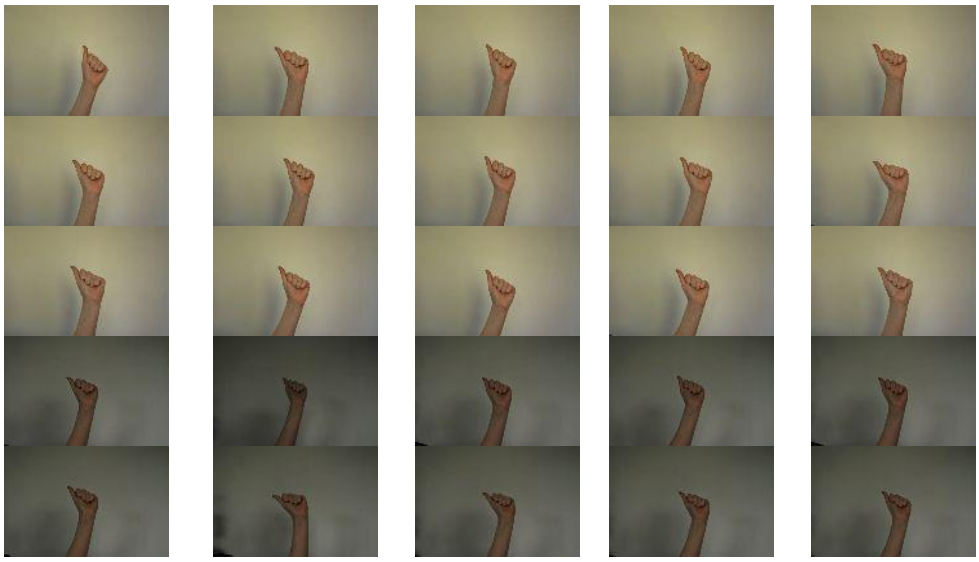


Figure 49 Set 1 class 1-a

Figure 50 Set 1 class 2-c

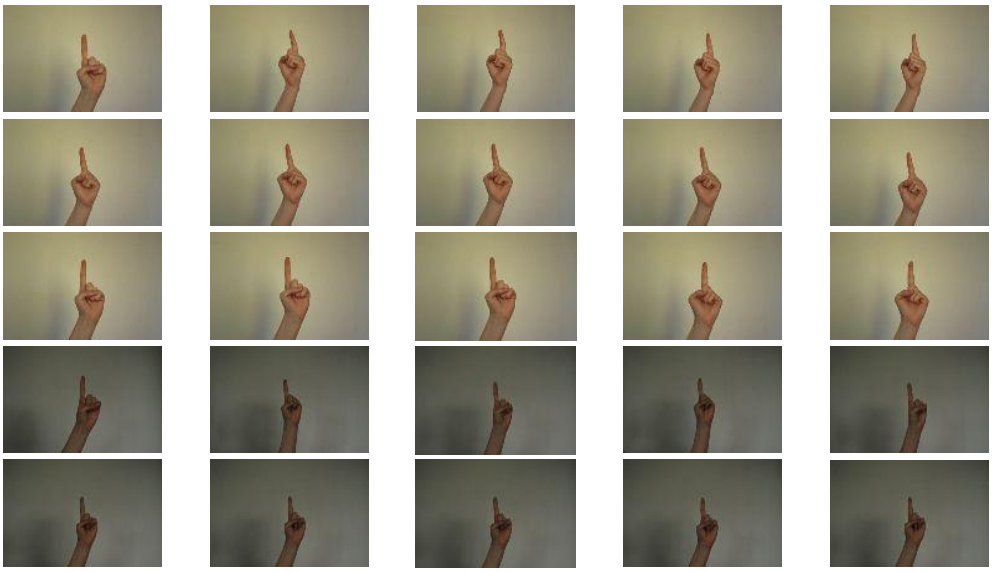


Figure 51 Set 1 class 3-d

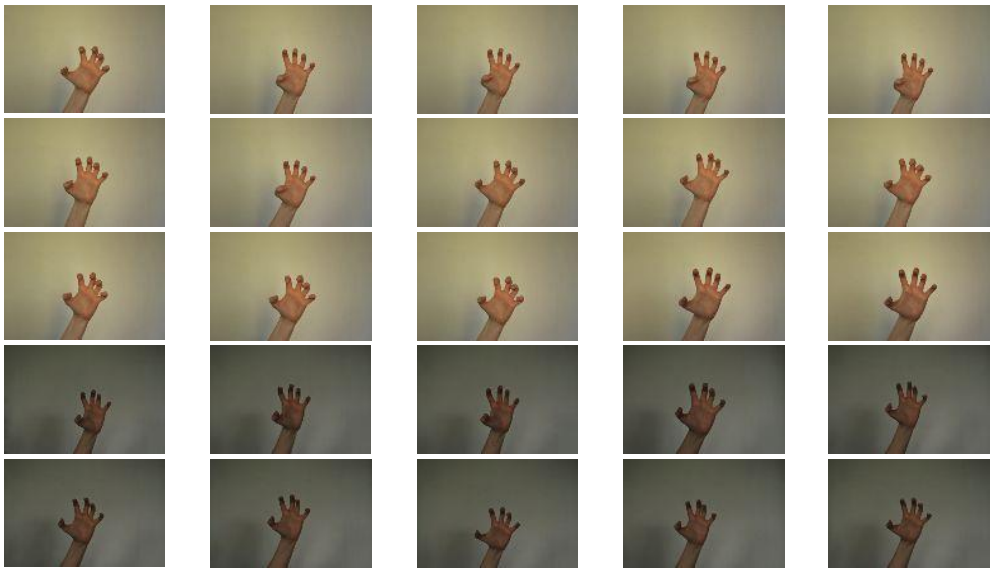


Figure 52 Set 1 class 4-e

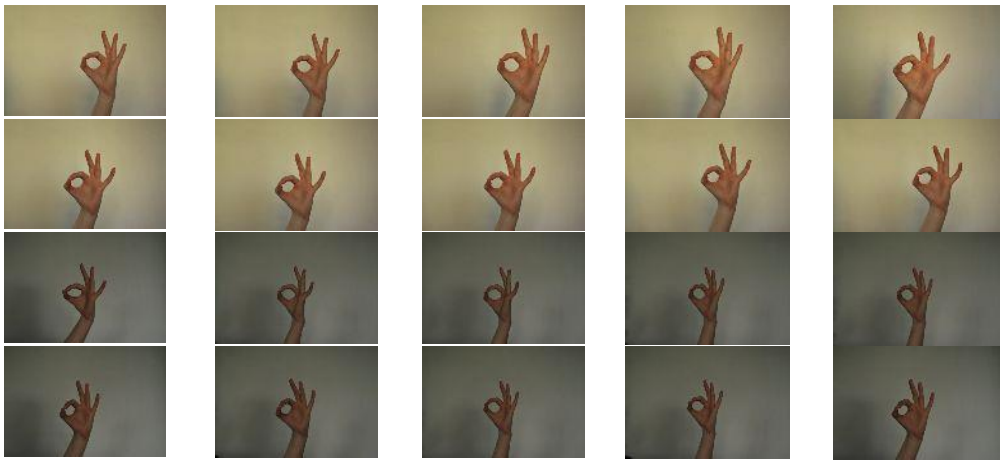


Figure 53 Set 1 class 5-f

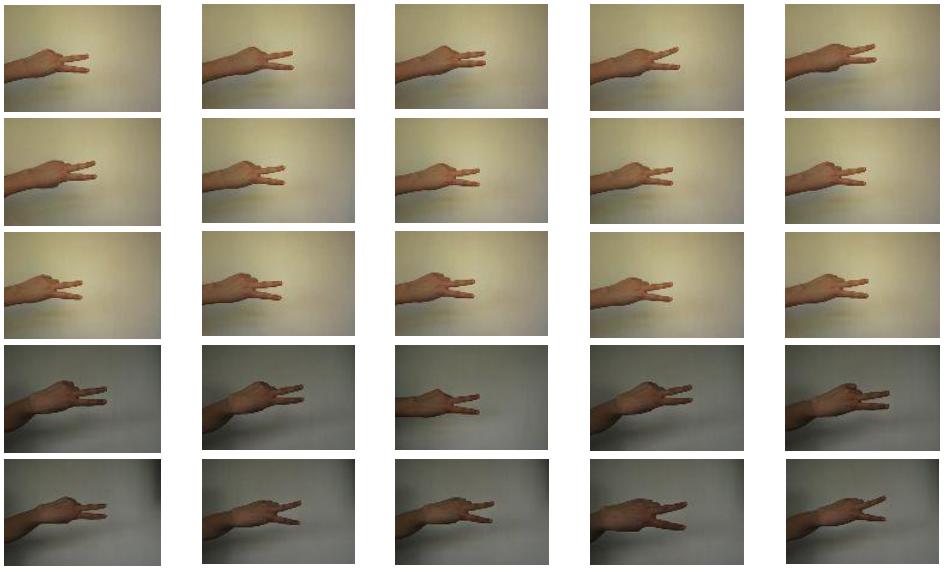


Figure 54 Set 1 class 6-p

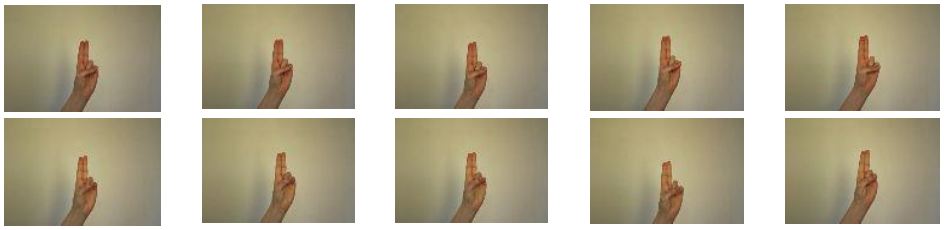


Figure 55 7Set 1 class 7-u

Figure 56 Set 1 class 8-w

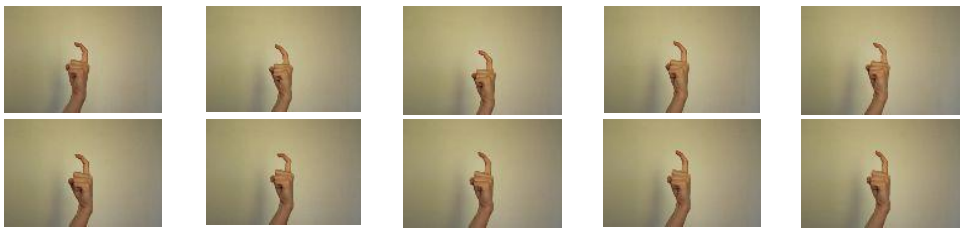


Figure 57 Set 1 class 9-x

In order to test the performances of the proposed method, because the number of samples per set is not very large the "leave one out" method is preferred.

5.2.1 Robust versus non-robust estimation of parameter λ

Due to the fact that parameter λ defines the balance between the two conflicting demands: generality and specificity its value proved to be very important in practice. Parameter λ reflects the way the generality and specificity combines in order to select the relevant prototypes which determinate farther the relevant composition used to describe an image.

Based on the proposed equation in [126] (see Eq. (5.2)), which is sensitive to noise and using the cost function,

$$S(\pi_i)\alpha - \log P(\pi_i) + \lambda H(P_{\pi_i}) \tag{5.1}$$

$$\lambda = \frac{\max_i \log P(\pi_i) - \min_i \log P(\pi_i)}{\max_i H(P_{\pi_i}) - \min_i H(P_{\pi_i})} \tag{5.2}$$

the following results are obtained: the error rate is 7.037% and confusion matrix can be seen in Table 1. The diagonal is 93.033%

Table 1 The Confusion Matrix for set 1 with 19 relevant composition , $\alpha = 0.02$, λ computed using Ommer [126] equation

Class / Predicted	a	c	d	e	f	p	u	w	x

a	90.1	3.3	3.3	0	0	0	0	0	3.3
c	3.3	93.4	0	0	0	0	0	0	3.3
d	0	0	100	0	0	0	0	0	0
e	3.3	0	0	93.4	3.3	0	0	0	0
f	0	0	3.3	0	86.8	0	0	0	9.9
p	0	0	0	0	0	100	0	0	0
u	0	0	6.6	0	0	0	93.4	0	0
w	0	0	0	0	0	0	0	100	0
x	3.3	9.9	6.6	0	0	0	0	0	80.2

Using the proposed robust estimation of the parameter λ (see Eq. (5.4)) and the cost function from Eq. (5.3):

$$S(\pi_i)\alpha - P(\pi_i) + \lambda H(P_{\pi_i}) \tag{5.3}$$

$$\lambda = \frac{IQR(P(\pi_i))}{IQR(H(P_{\pi_i}))} \tag{5.4}$$

The following results are obtained: the error rate is 3.70% and confusion matrix can be seen in Table 2. The diagonal is 96.29%.

Table 2 The Confusion Matrix for set 1 with 19 relevant composition prototypes, $\alpha = 0.02$, λ computed using the proposed equation

Class / Predicted	a	c	d	e	f	p	u	w	x
a	93.4	3.3	3.3	0	0	0	0	0	0
c	0	96.7	0	0	0	0	0	0	3.3
d	0	0	93.4	0	0	0	6.6	0	0
e	0	0	0	96.7	0	0	0	3.3	0
f	0	0	0	3.3	96.7	0	0	0	0
p	0	0	0	0	0	100	0	0	0
u	0	0	0	0	0	0	100	0	0
w	0	0	0	0	0	0	6.6	93.4	0
x	0	3.3	0	0	0	0	0	0	96.7

This results confirm the importance of the parameter λ. Using a robust estimation of this parameter the error rate decrees with 47.4%.

5.2.2 The importance of parameter α

As it was mention in chapter 4 parameter α rescale the relative coordinates of the compositions, because both relevant compositions and their position should have similar importance. A hand posture is recognized based on the relevant compositions and their relative position one to another.

Importance of parameter α is proved in the next experiments. For the same experiment like in 5.2.1 the value of parameter α is changed to 0.1. The results are summarized in Table 3. The diagonal is 93.4%, the error rate is 6.6%, and only one hand posture is recognized 100%.

Table 3 The Confusion Matrix for set 1 with 19 relevant composition prototypes, $\alpha = 0.1$, λ computed using the proposed equation

Class / Predicted	a	c	d	e	f	p	u	w	x
a	83.5	3.3	9.9	0	0	0	3.3	0	0
c	3.3	93.4	0	0	0	0	0	0	3.3
d	0	0	93.4	0	0	0	6.6	0	0
e	3.3	0	0	93.4	0	0	3.3	0	0
f	0	0	0	0	96.7	0	0	0	3.3
p	0	0	0	0	0	100	0	0	0
u	0	0	0	0	0	0	90.1	0	9.9
w	3.3	0	0	0	0	0	3.3	93.4	0
x	3.3	0	0	0	0	0	0	0	96.7

The results for $\alpha = 0.5$, 19 relevant composition prototypes, λ computed using the proposed equation can be seen in Table 4.

Table 4 The Confusion Matrix for set 1 with 19 relevant composition prototypes, $\alpha = 0.5$, λ computed using the proposed equation

Class / Predicted	a	c	d	e	f	p	u	w	x
a	80.2	3.3	13.2	0	0	0	0	0	3.3
c	6.6	90.1	0	0	0	0	0	0	3.3
d	0	0	96.7	0	0	0	3.3	0	0
e	0	0	0	93.4	0	0	3.3	0	3.3
f	0	0	3.3	0	90.1	0	3.3	0	3.3
p	0	0	0	0	0	100	0	0	0
u	3.3	0	3.3	0	0	0	86.8	0	6.6
w	3.3	0	0	0	0	0	9.9	86.8	0
x	6.6	0	3.3	0	0	0	0	0	90.1

The diagonal is 90.4%, the error rate is 9.6%, and only one hand posture is recognized 100%.

The results for $\alpha = 0.015$, 19 relevant composition prototypes, λ computed using the proposed equation can be seen in the confusion matrix from Table 5.

Table 5 The Confusion Matrix for set 1 with 19 relevant composition prototypes, $\alpha = 0.015$, λ computed using the proposed equation

Class / Predicted	a	c	d	e	f	p	u	w	x
a	93.4	3.3	3.3	0	0	0	0	0	0
c	0	93.4	0	0	0	0	0	0	6.6
d	0	0	93.4	0	0	0	6.6	0	0
e	0	0	0	93.4	0	0	0	6.6	0
f	0	0	0	3.3	96.7	0	0	0	0
p	0	0	0	0	0	100	0	0	0
u	0	0	0	0	0	0	100	0	0
w	0	0	0	0	0	0	6.6	93.4	0
x	0	3.3	0	0	0	0	0	0	96.7

The diagonal is 95.6% the error rate is 4.4%.

The results for $\alpha = 0.01$, 19 relevant composition prototypes, λ computed using the proposed equation can be seen in the confusion matrix from Table 6.

Table 6 The Confusion Matrix for set 1 with 19 relevant composition prototypes, $\alpha = 0.01$, λ computed using the proposed equation

Class / Predicted	a	c	d	e	f	p	u	w	x
a	93.4	3.3	3.3	0	0	0	0	0	0
c	0	93.4	0	0	0	0	0	0	6.6
d	0	0	93.4	0	0	0	6.6	0	0
e	0	0	0	93.4	0	0	0	6.6	0
f	0	0	0	3.3	96.7	0	0	0	0
p	0	0	0	0	0	100	0	0	0
u	0	0	0	0	0	0	93.4	0	6.6
w	0	0	0	0	0	0	6.6	93.4	0
x	0	3.3	0	0	0	0	0	0	96.7

The recognition rate is 94.8% the error rate is 5.2%.

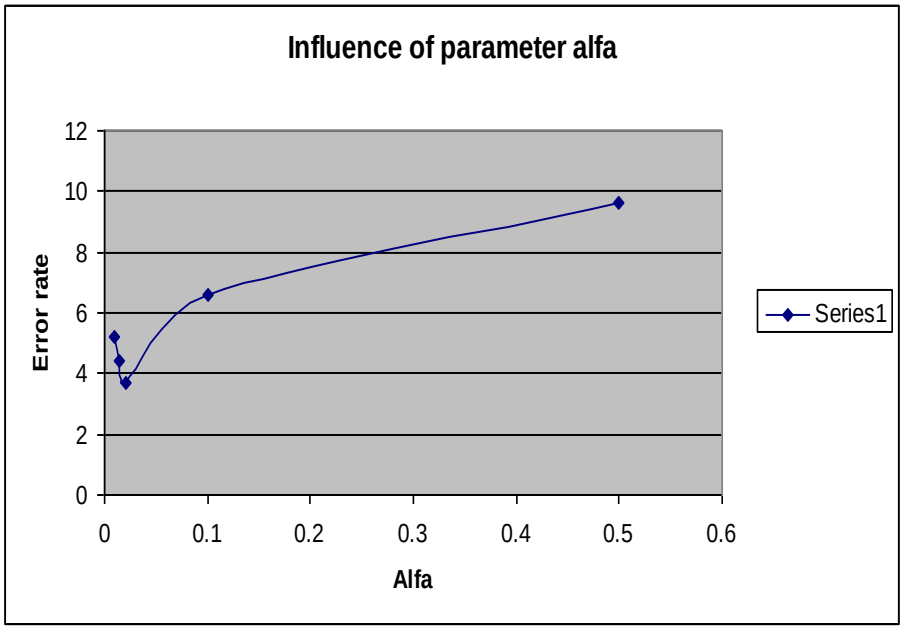


Figure 58 An illustration of the influence of parameter alfa.

These experiments prove the importance of parameter α, its best value is learned from the training data, and for our experiment it is 0.02.

5.2.3 Results regarding the clustering method

In order to know which clusering algoritm is better for the current data, 2 clustering algoritms are used. The k-means clustering algorithm is the one used to report the previous results. The scond cluseting algorithm is meanshift.The scale parameter for meanshift clustering algorithm is choose in order to have the same number of clusters *k,* and the same number of composition prototypes $\boldsymbol{\pi}_i \in \Pi$.

Using meanshift clustering algorithm 19 relevant composition prototype, the robust estimation of parameter λ and $\alpha = 0.02$ the error rate is much higher: 21.8%, the confusion matrix can be seen in Table 7.

Table 7 The Confusion Matrix for set 1 with 19 relevant composition prototypes, $\alpha = 0.02$, λ computed using the proposed equation, mean shift clustering algorithm

Class / Predicted	a	c	d	e	f	p	u	w	x
a	53.8	13.2	3.3	13.2	0	0	6.6	0	9.9
c	0	57.1	0	6.6	0	0	19.8	9.9	6.6
d	0	0	86.6	0	0	0	3.3	0	9.9
e	3.3	0	0	96.7	0	0	0	0	0
f	3.3	3.3	0	13.2	63.7	0	9.9	3.3	3.3
p	0	0	0	0	0	100	0	0	0
u	0	0	6.6	0	0	0	83.5	0	9.9
w	3.3	0	0	0	0	0	3.3	93.4	0
x	0	3.3	0	0	0	0	26.4	0	96.7

The k-means clustering algorithm has very good results when it deals with spherical clusters.

5.2.4 Experiments for different numbers of relevant composition prototypes.

For 14 relevant compositon prototypes, keeping all the other parameters at the same value $\alpha = 0.02$, k-means clustering, the error rate is 70.2% and the recognition rate is 29.8%

Table 8 The Confusion Matrix for set 1 with 14 relevant composition prototypes, $\alpha = 0.02$, λ computed using the proposed equation, k-means clustering algorithm

Class / Predicted	a	c	d	e	f	p	u	w	x
a	30.7	0	49.5	3.3	0	0	0	13.2	3.3
c	19.8	10.9	16.5	23.1	9.9	0	0	3.3	16.5
d	6.6	0	83.5	3.3	0	0	0	0	6.6
e	9.9	0	0	27.4	3.3	0	0	29.7	0
f	0	0	49.5	26.4	14.2	0	0	9.9	0
p	9.9	9.9	3.3	23.1	0	40.6	0	0	13.2
u	6.6	0	46.2	0	0	0	7.6	0	39.6
w	9.9	0	26.4	0	9.9	0	0	53.8	0
x	16.5	0	75.9	3.3	0	0	0	3.3	0

For 16 relevant compositon prototypes, keeping all the others parameters at the same value $\alpha = 0.02$, k-means clustering, the error rate is 7%, the diagonal is 93% as it can be seen in Table 9.

Table 9 The Confusion Matrix for set 1 with 16 relevant composition prototypes, $\alpha = 0.02$, λ computed using the proposed equation, k-means clustering algorithm

Class / Predicted	a	c	d	e	f	p	u	w	x
a	86.8	3.3	9.9	0	0	0	0	0	0
c	0	86.8	0	0	0	0	0	0	13.2
d	0	0	100	0	0	0	0	0	0
e	3.3	0	0	93.4	3.3	0	0	0	0
f	0	3.3	0	6.6	90.1	0	0	0	0
p	0	0	0	0	0	100	0	0	0
u	0	0	9.9	0	0	0	86.8	0	3.3
w	0	0	0	0	0	0	0	100	0
x	0	3.3	0	0	0	0	3.3	0	93.4

For 18 relevant compositon prototypes, keeping all the others parameters at the same value $\alpha = 0.02$, k-means clustering, the error rate is 4.4%, the diagonal is 95.6 %. The confusion matrix for this experiment can be seen in Table 10.

Table 10 The Confusion Matrix for set 1 with 18 relevant composition prototypes, $\alpha = 0.02$, λ computed using the proposed equation, k-means clustering algorithm

Class / Predicted	a	c	d	e	f	p	u	w	x
a	90.1	3.3	6.6	0	0	0	0	0	0
c	0	96.7	0	0	0	0	0	0	3.3
d	0	0	100	0	0	0	0	0	0
e	3.3	0	0	93.4	3.3	0	0	0	0
f	0	0	0	3.3	96.7	0	0	0	0
p	0	0	0	0	0	100	0	0	0
u	0	0	6.6	0	0	0	90.1	0	3.3
w	0	0	0	0	0	0	0	100	0
x	0	3.3	0	0	0	0	3.3	0	93.4

In Figure 59 and Figure 60 the evolution of recognition rate per class for diffrent numbers of relevant composition prototypes and the evolution of error rate and recognition rate for r=14, 16, 18 and 19 is shown.

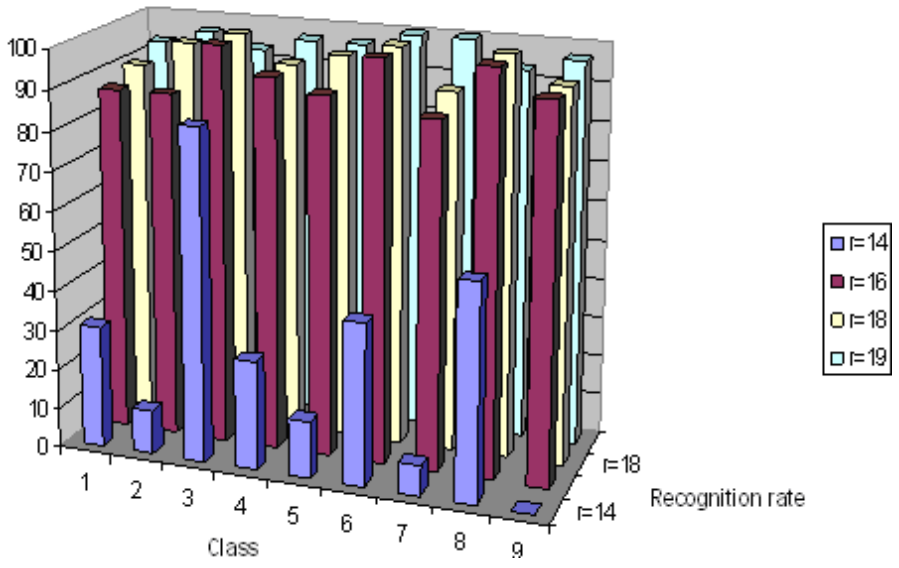


Figure 59 Recognition rate per class for diffrent numbers of relevant composition prototypes

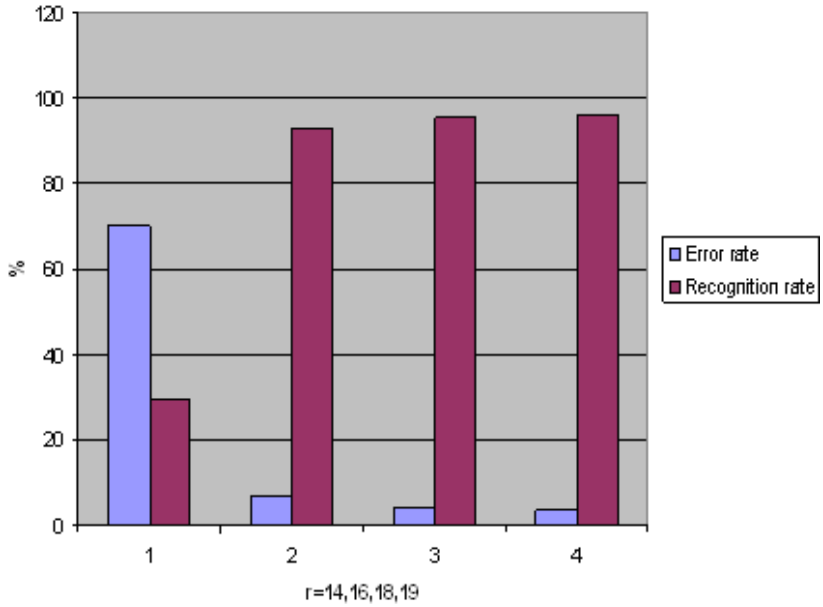


Figure 60 The evolution of error rate and recognition rate for r=16,18 and 19

5.3 Experiments results for set 2

The second set of hand postures has six classes. This six had postures are chosen by the considerate that they are easy to perform in front of a webcam by a person while is sited. The pictures from set 2 are taken in different days, with different day light illuminations and fluorescent light. The background is a white paper and a brown carpet. The training set has 60 samples per class and the testing sets have other 30 training samples per class. These images are acquired by a Canyon webcam- CN-WCAMNI.

For set 2 the images resolution is 640px x 480px.

In order to extract the hand contours Canny edge detector from Image Processing Toolbox, Matlab is used. The sensitivity thresholds for the Canny method for set 1 is defined; the high threshold *thresh* is 0.4 and 0.4**thresh* is used for the low threshold.

The two bin colour histogram is extract only for the red component of the RGB image. Parameter σ from Eq.(4.6) is equal to 1 and parameter σ from Eq.(4.11) equal to 0.05.

The number of composition prototypes is 30 and the number of relevant composition prototypes *r*, which conduct to the best result is equal to 28.

The training set has as background a white paper as it can be seen in figures below. No artificial light was added to the training images.

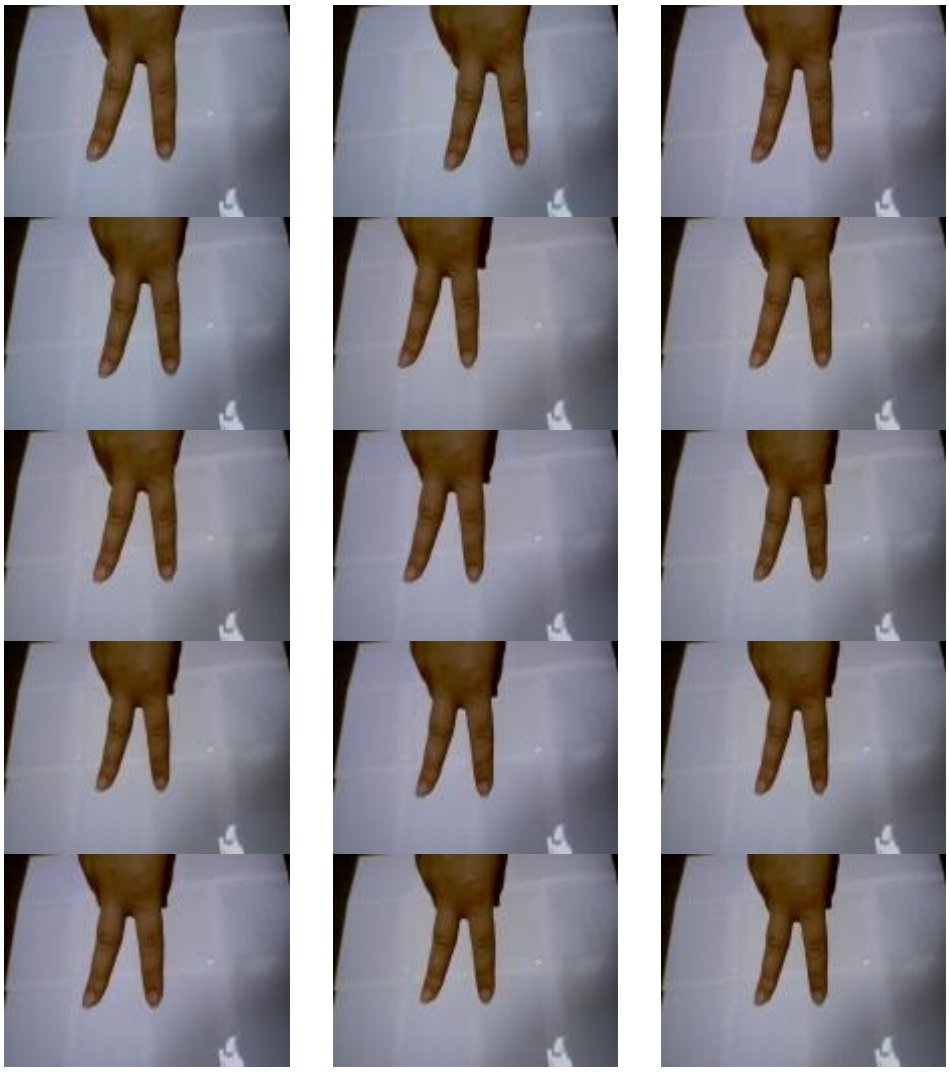


Figure 61 Images from training set 2- class 1

Figure 62 Images from training set 2- class 2

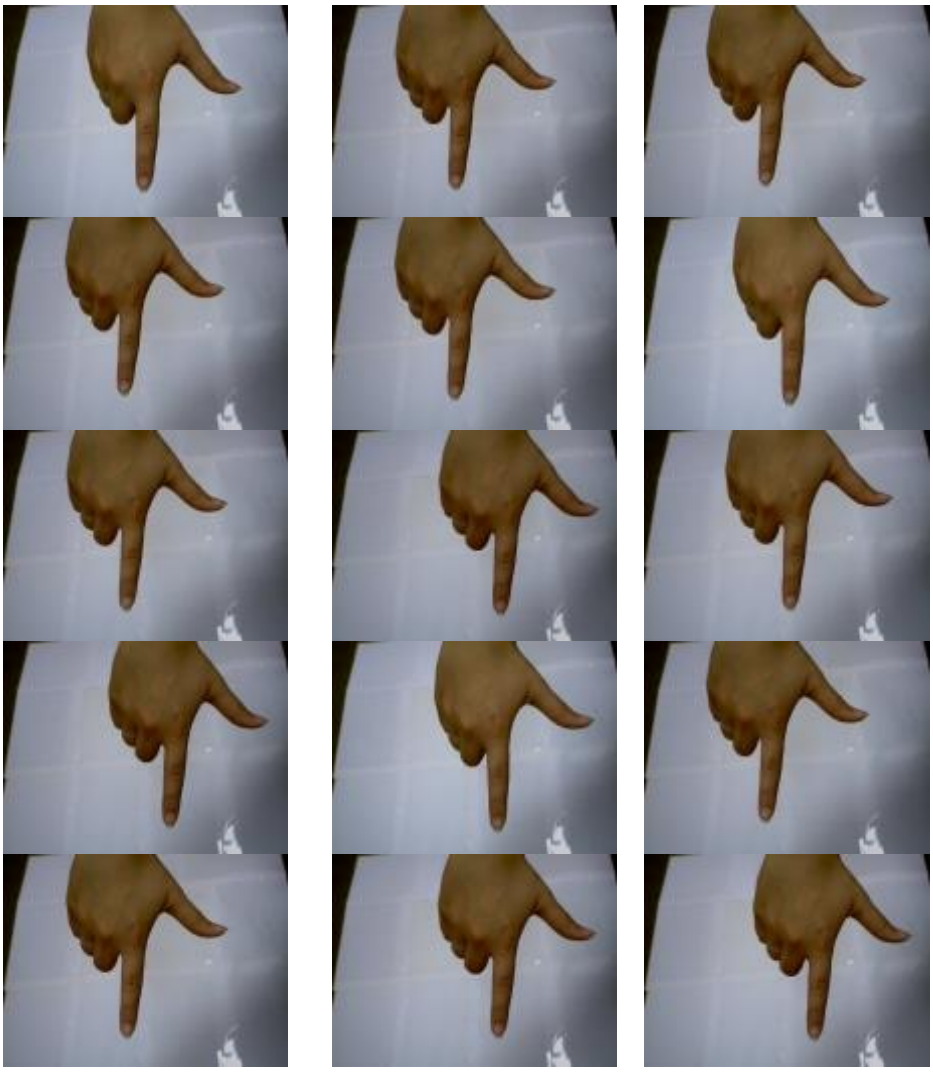


Figure 63 Images from training set 2- class 3

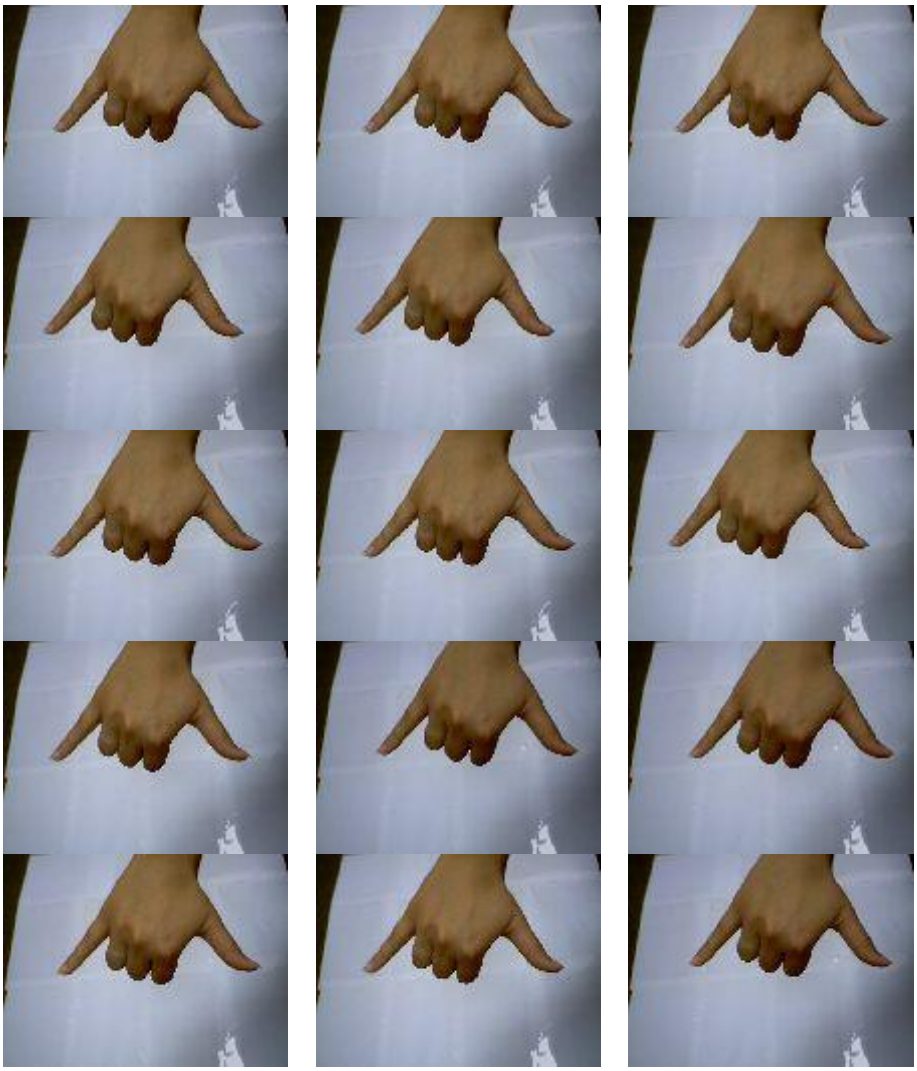


Figure 64 Images from training set 2- class 4

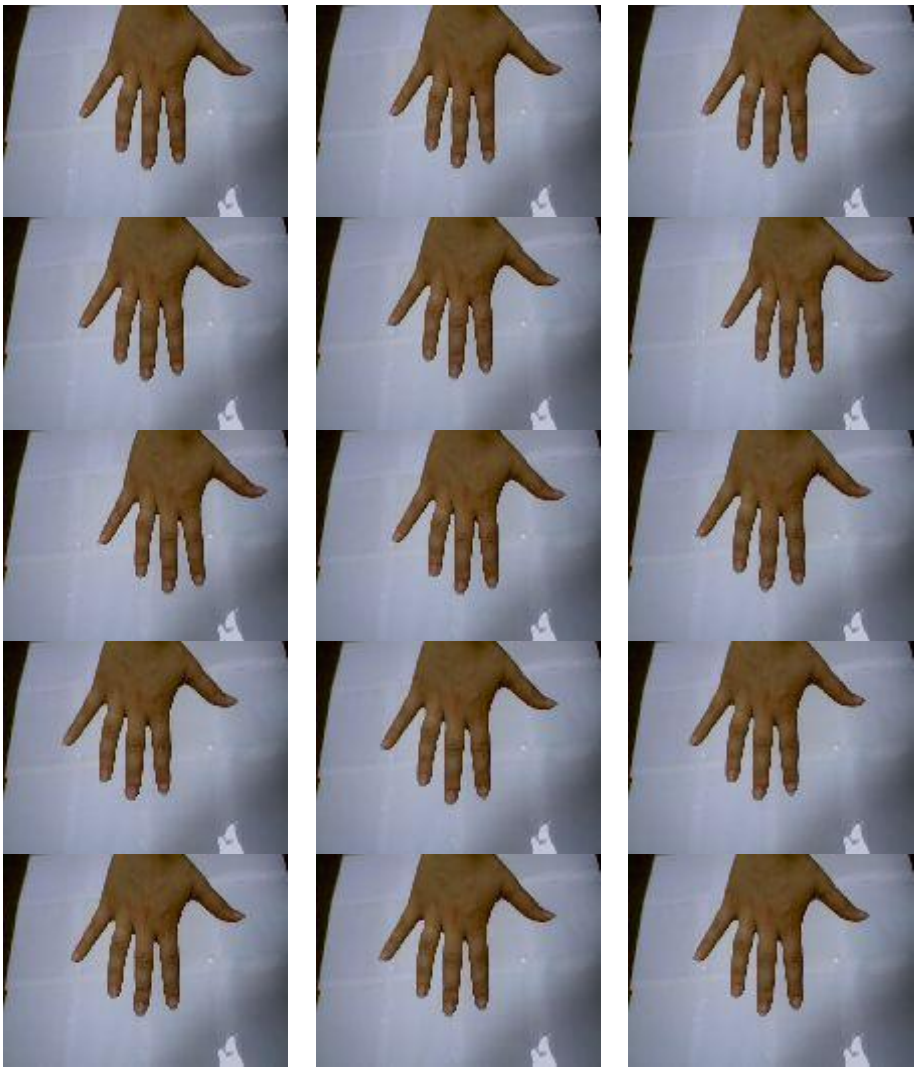


Figure 65 Images from training set 2- class 5

Figure 66 Images from training set 2- class 6

5.3.1 Experiments using " leave one out" method

For the first experiments „leave one out" method is uesd.

Using Ommer [126] estimation of the parameter λ and the cost function from Eq.(5.1) the following results summerized in Table 11 are obtained.

Table 11 The Confusion Matrix for set 2 with 28 relevant composition prototypes, $\alpha = 0.02$, λ estimated using Ommer equation, k-means clustering algorithm

Class / Predicted	1	2	3	4	5	6
1	100	0	0	0	0	0
2	0	99.72	0.28	0	0	0
3	0	0	99.45	0	0.55	0
4	0	0	0	100	0	0
5	0	0	0.55	0	99.45	0
6	0	0	0.55	0	0.55	98.9

The recognition rate is 99.59% and the error rate is 0.41%.

Using the proposed equation for estimation of the parameter λ and the cost function from Eq.(5.3) the results from Table 12 are obtained

Table 12 The Confusion Matrix for set 2 with 28 relevant composition prototypes, $\alpha = 0.02$, λ computed using the proposed equation, k-means clustering algorithm

Class / Predicted	1	2	3	4	5	6
1	100	0	0	0	0	0
2	0	100	0	0	0	0
3	0	0	100	0	0	0
4	0	0	0	100	0	0
5	0	0	0.55	0	99.45	0
6	0	0	0.55	0	0	99.45

For this case the average recognition rate is 99.82% and the error rate is 0.18%.

5.3.2 Experiments with new test images

In the next experiments there are 30 new samples per class and the background is a white paper. In figures 67 to 72, these new testing images can be seen.

The results obtain with the non-robust estimation of the parameter λ are presented in the confusion matrix from Table 13.

Table 13 The Confusion Matrix for set 2 with 28 relevant composition prototypes, $\alpha = 0.02$, λ computed using Ommer equation, k-means clustering algorithm

Class / Predicted	1	2	3	4	5	6
1	100	0	0	0	0	0
2	3.3	96.7	0	0	0	0
3	0	0	93.4	0	3.3	3.3
4	0	0	0	100	0	0
5	0	0	0	0	100	0
6	9.9	0	0	0	3.3	86.8

The recognition rate is 96.15% and the error rate is 3.75%.

The results for the same test images but with λ computed with the proposed equation are presented in Table 14.

Table 14 The Confusion Matrix for set 2 with 28 relevant composition prototypes, $\alpha = 0.02$, λ computed using the proposed equation, k-means clustering algorithm

Class / Predicted	1	2	3	4	5	6
1	100	0	0	0	0	0
2	3.3	96.7	0	0	0	0
3	0	0	93.4	0	3.3	3.3
4	0	0	0	100	0	0
5	0	0	0	0	100	0
6	3.3	0	0	0	3.3	93.4

The recognition rate is 97.25% and the error rate is 2.75%.

Figure 67 Images from testing set 1-class 1

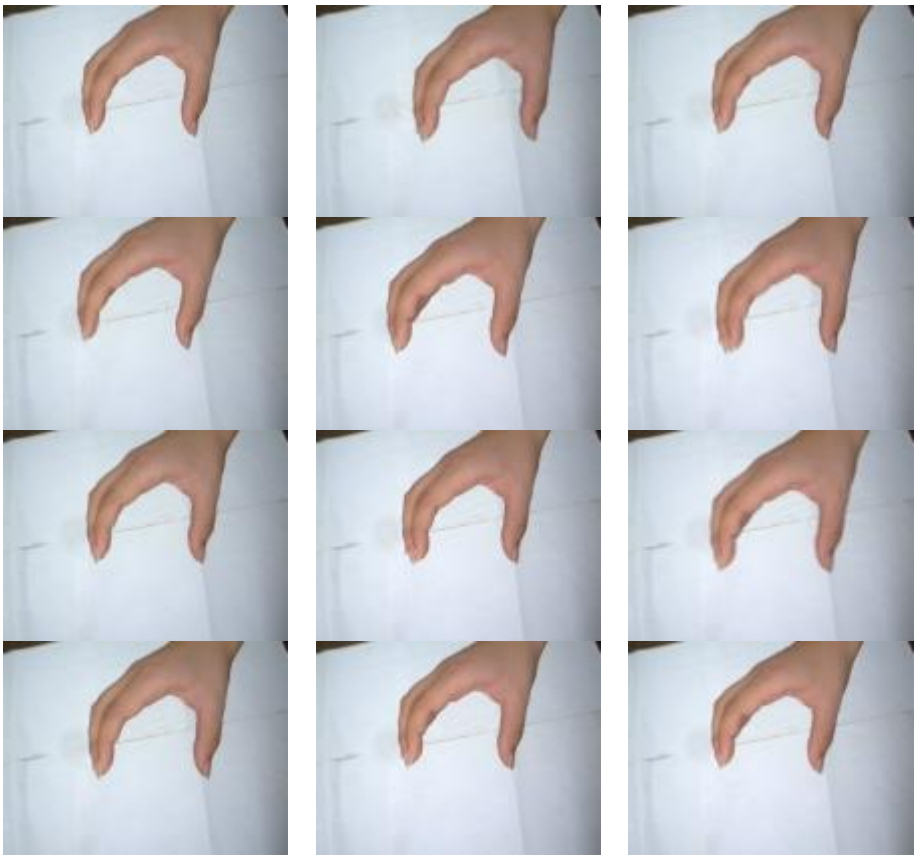


Figure 68 Images from testing set 1-class 2

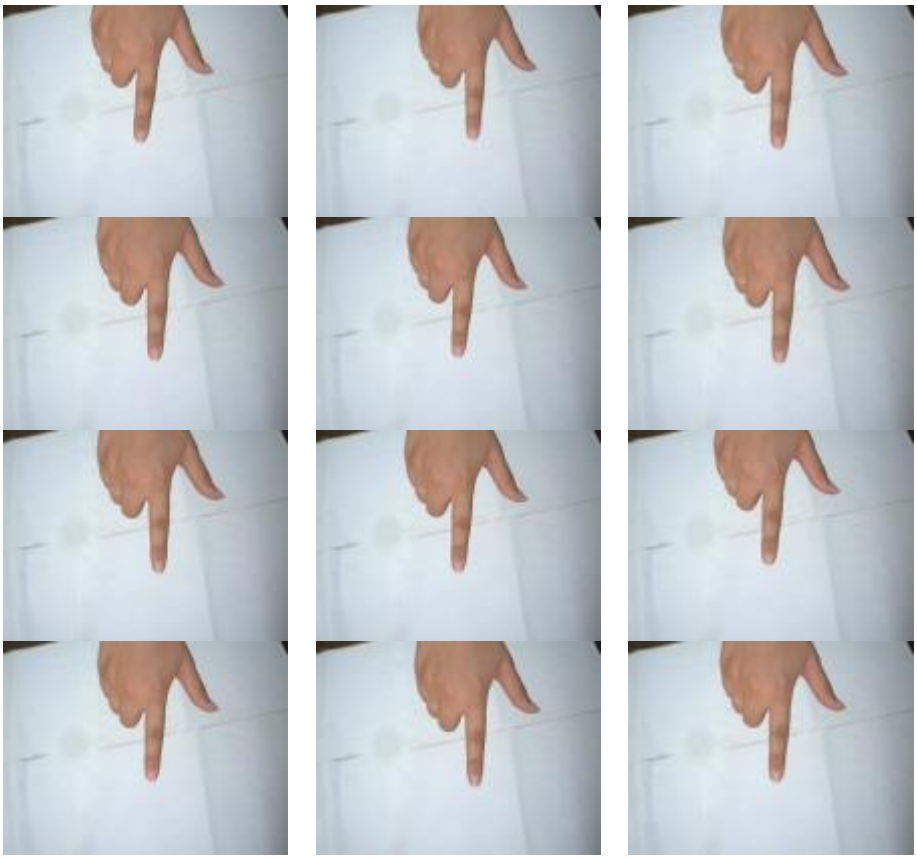


Figure 69 Images from testing set 1-class 3

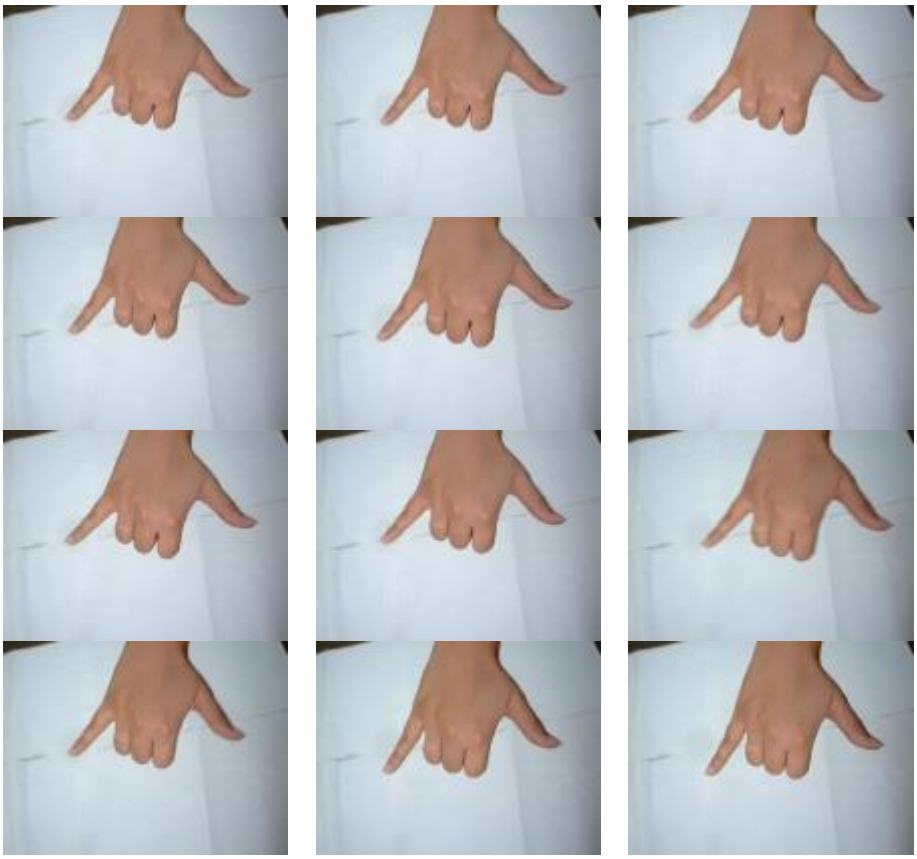


Figure 70 Images from testing set 1-class 4

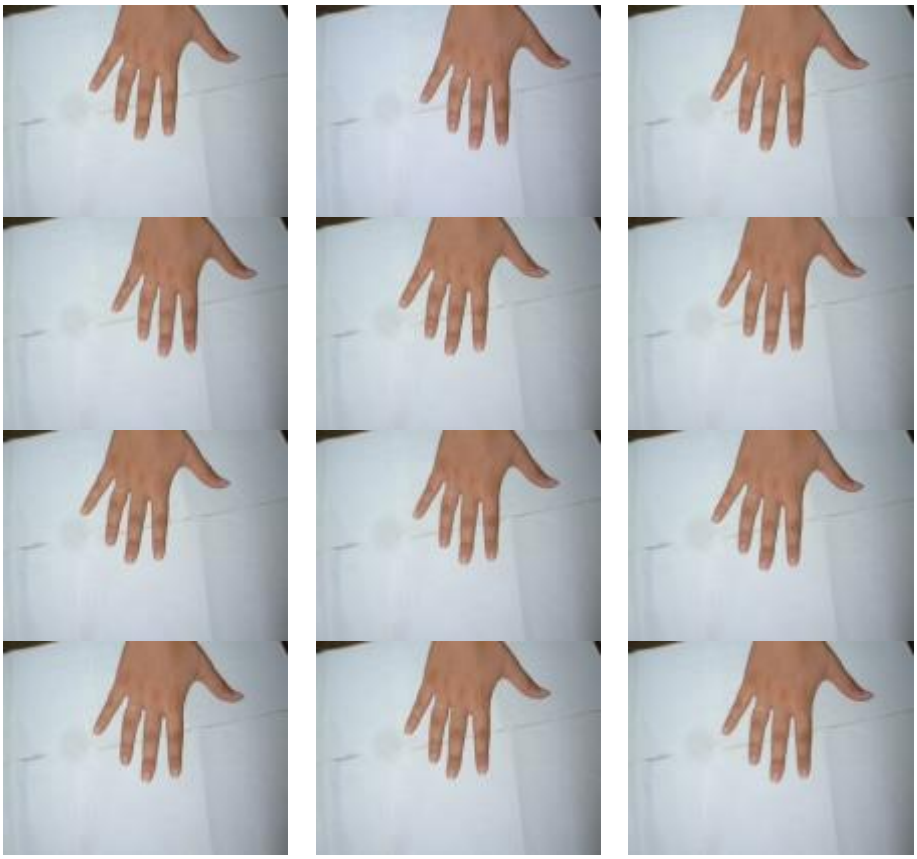


Figure 71 Images from testing set 1-class 5

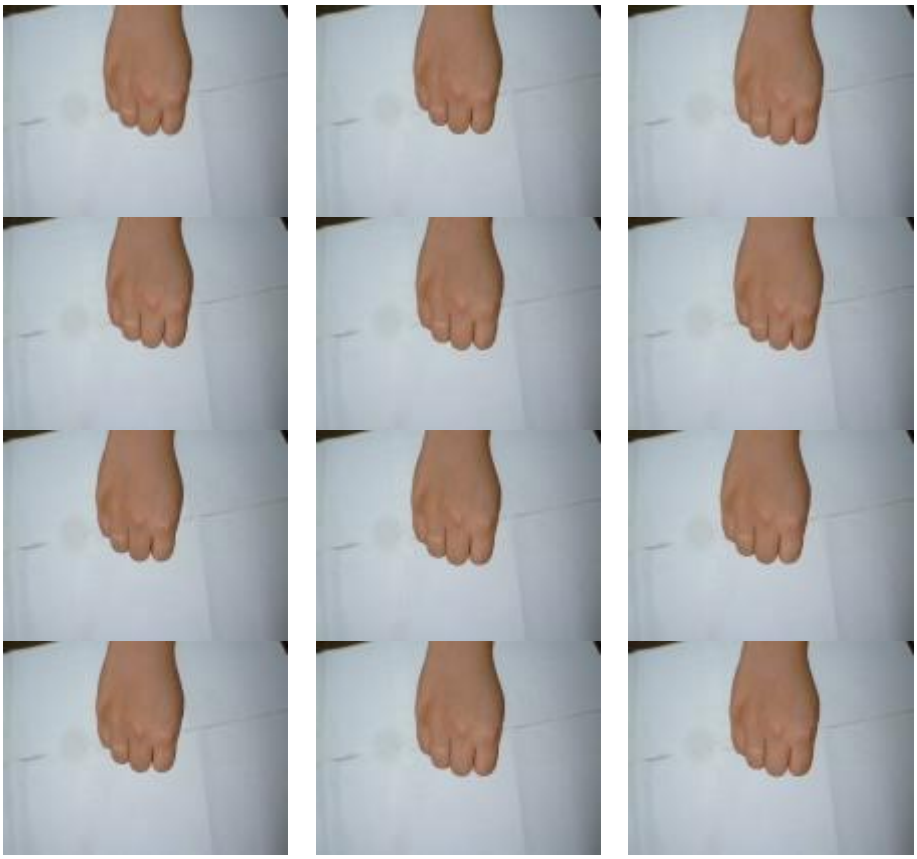


Figure 72 Images from testing set 1-class 6

In the next experiments there are 30 new samples per class and the background is a brown carpet. In figures 73 to 78, some of these new testing images can be seen.

For this set of images the following results summarized in Table 15 are obtained when the parameter λ was computed with Ommer equation.

Table 15 The Confusion Matrix for set training set 2 with 28 relevant composition prototypes, $\alpha = 0.02$, λ computed using Ommer equation, k-means clustering algorithm

Class / Predicted	1	2	3	4	5	6
1	94.6	0	0	0	3.3	0
2	6.6	96.7	0	0	0	0
3	0	0	93.4	0	3.3	3.3
4	0	0	0	93.4	0	6.6
5	0	0	0	0	100	0
6	6.6	0	0	0	3.3	90.1

The error rate for this case is 94.7% and error rate is 5.3%.

When parameter λ was computed with the proposed equation, the experiments conducted to the following results presented in the matrix confusion from Table 16.

Table 16 The Confusion Matrix for set training set 2 with 28 relevant composition prototypes, $\alpha = 0.02$, λ computed using the proposed equation, k-means clustering algorithm

Class / Predicted	1	2	3	4	5	6
1	100	0	0	0	0	0
2	3.3	96.7	0	0	0	0
3	0	0	93.4	0	3.3	3.3
4	0	0	0	93.4	0	6.6
5	0	0	0	0	100	0
6	6.6	0	0	0	3.3	90.1

Recognition rate is 95.6% and error rate is 4.4%

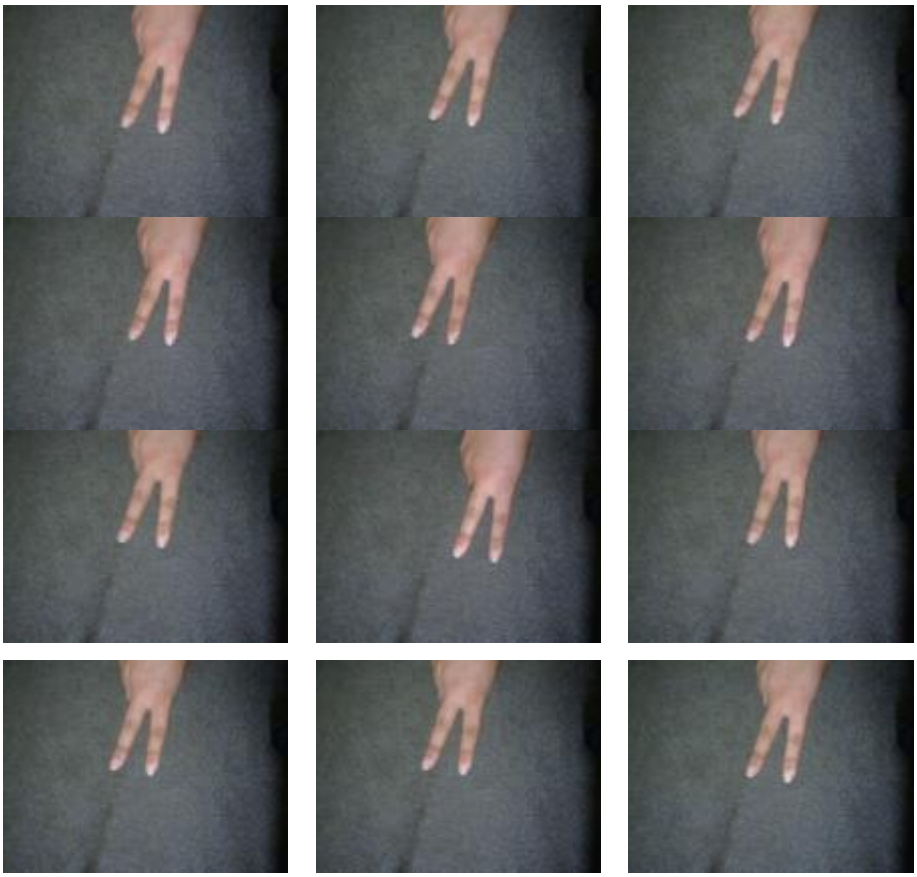


Figure 73 Images from testing set 2-class 1

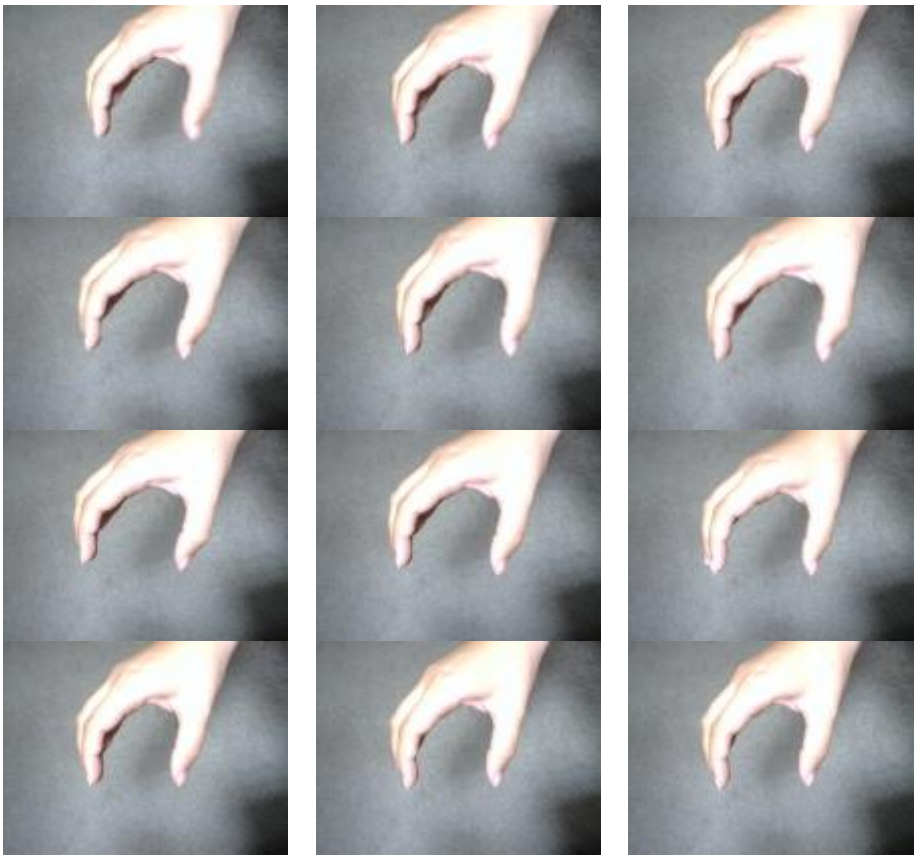


Figure 74 Images from testing set 2-class 2

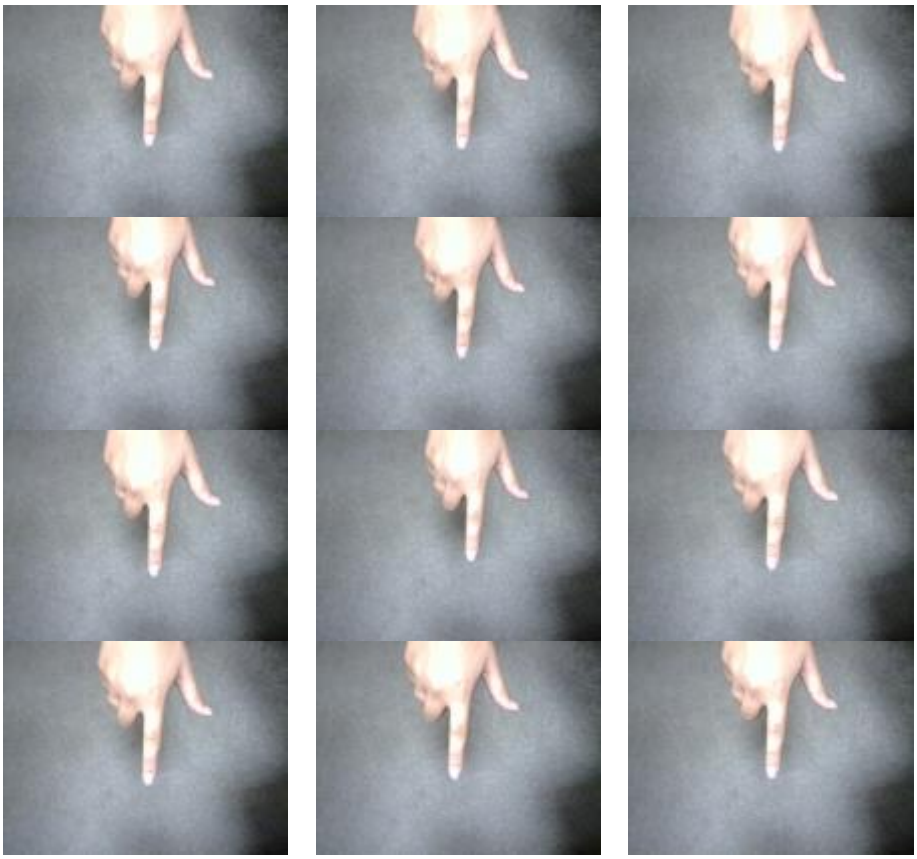


Figure 75 Images from testing set 2-class 3

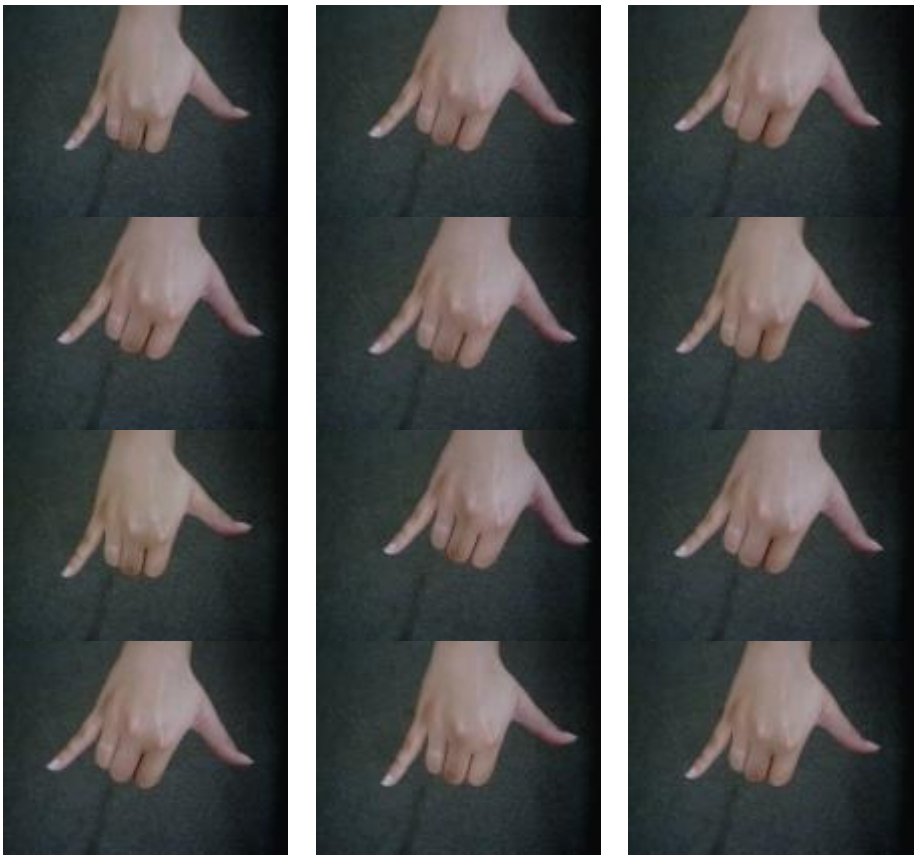


Figure 76 Images from testing set 2- class 4

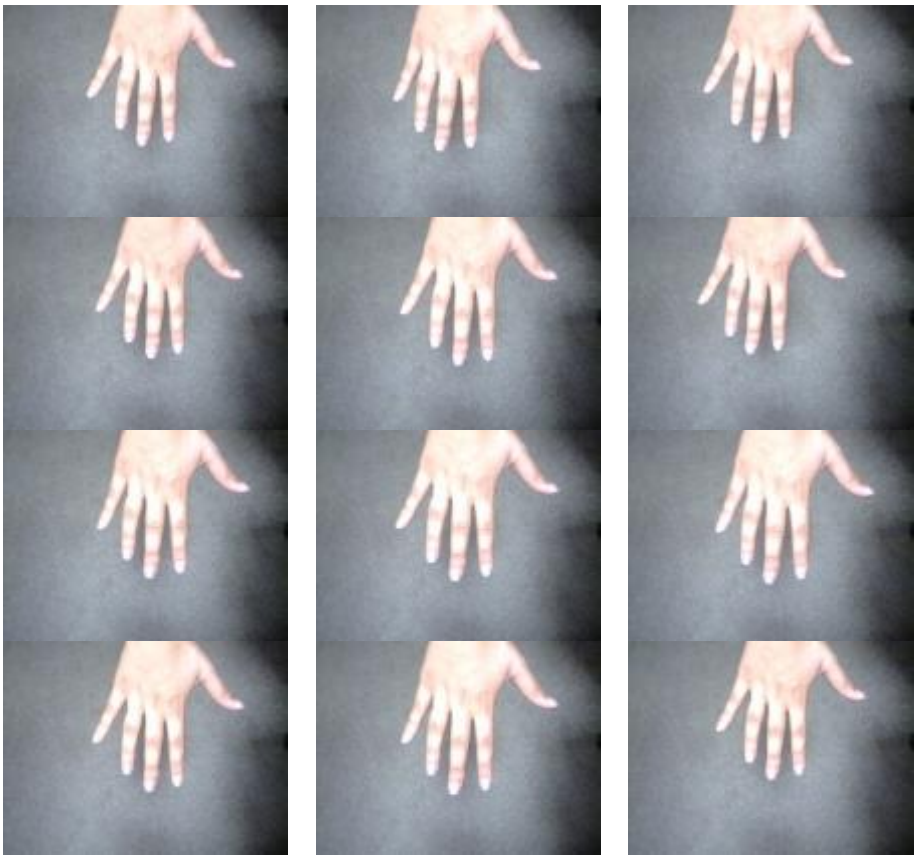


Figure 77 Images from testing set 2- class 5

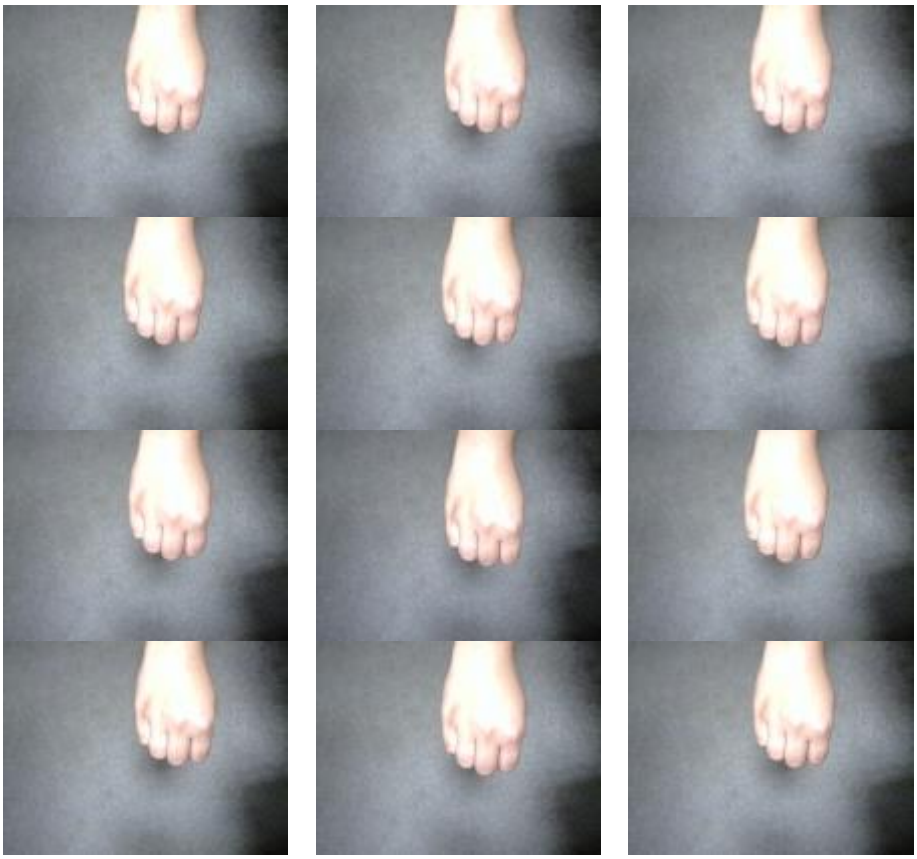


Figure 78 Images from testing set 2- class 6

5.4 Conclusions

In this chapter the experimental results which prove the potential of the compositional tehniques are presented. Our best result for the first set of images which consists of nine classes is 96.29%; the best result for the secod set of images which consists of 6 clases is 99.82%.

The experiments also prove the importance of parameter λ, which makes a trade-off between general and specific in the cost function defined in Eq. (4.15). The robust estimation of parameter λ in order to select the relevant compositions prototypes represents a major asset of this work. Using the robust estimation of parameter λ for set 1 of images, the recognition rate was 96.29% and using the non-robust estimation of the parameter λ, the error rate for the same experiment was 93.033%. For set 2 of images the recognition rate was 96.82% when „leave one out" method was used. For the same experiment using the non-robust estimation of parameter λ the recogniton rate was 99.59%. The results obtained for a new set of hand postures (different from the training images) are: 97.25% when we used the value of parameter λ estimated with the proposed Eq.(4.18) and 96.15% when its value was estimated using Ommer equation. For the second set of test images the background was a brown carpet and the following results we had: for the robust estimation of parameter λ the recognition rate was 95.6% and for the non-robust estimation of parameter λ the recognition rate was 94.7%.

Based on relevant compositions prototypes the relevant compositions are selected. Relevant compositions and their rescaled position is used to describe the image. Both relevant compositions and their position should have similar importance because the hand posture is recognized based on types of compositions and their relative position one to another. In order to have this, the parameter α is introduced and its value is learned based on the experimental data. The importance of parameter α is shown in experiments, the best recognition rate 96.29% was obtained for α =0.02.

The number of relevant composition prototypes proves to have a grate influence in practice. For the first set of hand postures the best recognition rate 96.29%, was obtained for 19 relevant composition prototypes. For 14 relevant composition prototypes the recognition rate has decreased dramatically to 29.8%; for 16 relevant composition prototypes the recognition rate was 93%, and for 18 relevant composition prototypes it was 95.6%.

This results compare favorably with other results reported in literature as it can be seen in Table 17.

Table 17 Results for hand posture recognition

No. of hand postures	No. of test images	Recognition rate	Source
3	275	95.6%	Wang C .C, Wang K. C., [211], 2008
4	400	95.75%	Chen Q., Georganas N.D., Petriu E.M, [212] 2007
6	**360**	**99.82%**	**Current work**
9	**270**	**96.29%**	**Simion G.,Gui V.,OtesteanuM. [9] 2009.**
10	718	92.79%	Just, A. Rodriguez, Y. Marcel, S. [213] 2006.

It is quite difficult to compare our results with those reported in literature, because there are diffrent data bases, with diffrent hand potures and a diffrent number of hand postures. However the results obtained witht the proposed compositional method show that compositional methods should be rekon in object recognition and more specific in hand posture recognition.

6 CONCLUSIONS

6.1 Discussions

This thesis addresses to hand posture recognition. Hand posture recognition is an important task for Computer Vision researchers, because nowadays the use of hand gestures has become an important part of human computer interaction (HCI). To use human hands as a natural interface device, some glove-based devices have been employed to capture human hand motion by attaching sensors to measure the joint angles and spatial positions of hands directly. These devices have the disadvantages of being expensive and cumbersome. Vision-based techniques provide promising alternatives to capture human hand motion being cost efficient and noninvasive. These facts serve as the motivating forces for research in the modeling, analysis, animation, and recognition of hand gestures.

There is a large variety of applications which involves hand gestures. Hand gestures can be used to achieve natural human computer interaction for virtual environments; there are attractive methods for communication with the deaf and dumb. An important application area is that of vehicle interfaces. The primary motivation of research into the use of hand gestures for in-vehicle secondary controls is broadly based on the premise that taking the eyes off the road to operate conventional secondary controls can be reduced by using hand gestures. The healthcare area also could benefit by the hand gesture recognition systems. The gesture based system could replace touch screens now used in many hospital operating rooms which must be sealed to prevent accumulation or spreading of contaminants and requires smooth surfaces that must be thoroughly cleaned after each procedure.

This thesis proposes a compositional approach to hand gesture recognition. The main advantage of the compositional techniques is their generality; these techniques are more independent of application. Using these techniques we address also to the semantic gap that exists between the low level features and high level representations. Using the compositional method characteristic regions from hand are extracted. The hand posture is no longer modeled as a whole. This characteristic regions are assembled to form compositions, this compositions at their turn can be group in compositions of compositions and so on. Using these methods partial occlusions of object can be handle. These methods allow us to incorporate the Gestalt laws of visual perception. These laws are a set of visual rules that guide the construction process of groupings and yield compositions, establishing causal relationships between grouping constituents, and tends to emulate the way our brain-view processor works better.

6.2 Contributions

The contributions of this work are:

- **the compositional approach used to hand posture recognition**.

According to Principle of Compositionality, we tend to decompose object in simple parts, and represent the object as hierarchies of meaningful parts. Following this principle a hand gesture is decompose in: the V shapes between the fingers when these are apart, the curve shapes which correspond to the fingertips and the straight lines from the finger length. Each hand pose can be defined as a combination of these shapes. Based on the number of V shapes, curves and lines and based on the relations among them the hand pose can be recognized.

- **careful selection of the basic features**

Contours, interest points, patchs, colour histograms, orientation histograms, are used to generate the V shapes, curves and lines. Another contribution of this work consists in the **selection of the V shape, curves and lines** which are used to describe a hand posture.

- **the choice of the right distance between the parts which are about to be grouped.**

The object representation is based on *compositions* of parts: descriptors are grouped according to Gestalt law of proximity to obtain a set of possible candidate compositions. In order to generate the compositions of parts it was important to choose the right distance between the parts which are about to be grouped. *Candidate* compositions from all test images are clustered and the resulted composition prototypes are used to form the composition codebook. Based on the cost function the relevant compositions prototypes are learned in the next stage.

- **the optimization of parameter** λ

The robust estimation of parameter λ in order to select the relevant compositions prototypes represents a major asset of this work.

- **the use of parameter** α

Based on relevant compositions prototypes the relevant compositions are selected. Relevant compositions and their rescaled position are used to describe the image. Both relevant compositions and their position should have similar importance because the hand posture is recognized based on types of compositions and their relative position one to another. In order to have this, the parameter α is introduced and its value is learned based on the experimental data.

- **the proposed discriminant function**

The proposed discriminant function used in classification was inspired by the point matching used in image registration.

- **the development of the software package** used to recognize hand postures

- **the development of a data base for hand posters** is also a contribution of this thesis.

6.3 Further work

This work proves the potential of compositional techniques for hand gesture recognition. The study of compositional techniques has gained popularity in the last years, so for the future there are a lot of things to be explored.

Regarding that work, different types of optimizations can be tried: from patch size to law of groping that is used to grouped parts in order to form compositions. Different feature detectors and feature descriptors could be tested. It is also possible to introduce new intermediate layers of abstraction between feature extraction and classification.

In the future the algorithm has to be tested on different subjects.

This method is quite general so in the future we plan to use it to recognize different objects, not only hand postures.

BIBLIOGAPHY

[1] Rugină S., **Sârbu G**., Otesteanu M., Gontean A. "Analysis and modeling of rain characteristics." *Proceedings of the Symposium on Electronics and Telecommunications ETC'2006,* 51 (2) (2006): 215-18

[2] Gontean A., Otesteanu M., Rugină S., **Sârbu G**. "Versatile Communication Solution for PLC Based Control Systems." *WSEAS Transactions on Communications,* 12 (5) (2006): 2137-41

[3] Otesteanu M., Gontean A., **Sârbu G**., Rugină S. "Software Environment for the Laser Precipitation Monitor." *WSEAS Transactions on Information Science and Applications* 4 (1) (2007): 214-19

[4] **Sârbu G**., Rugină S., Otesteanu M., Gontean A. "Target Detection Using A Laser Sensor." *MicroCAD 2007, International Scientific Conference, Section I: Automation and Telecommunication* (2007): 23-28

[5] **Simion G**., Otesteanu M., Gontean A. "Target Detection in Low Visibility Condition and artificial Lighting Using a Laser Sensor." *Computational Intelligence Man-Machine Systems and Cybernetics* (2007)

[6] **Simion G.**, Gui V., Otesteanu M., Popa D., David C. "Hand Edge Detection for Gesture Analysis in a Sparse Framework." *Buletinul stiintific al Universitatii Politehnica din Timisoara* 53 (2) (2008): 155-60

[7] **Simion G.,** Gui V., Otesteanu M. "Hand Posture Recognition Using Compositional Techniques." *5th International Symposium on Applied Computational Intelligence and Informatics (SACI)* (2009): 435-39

[8] **Simion G**., Gui V., Otesteanu M. "A New Compositional Technique for Hand Posture Recognition." *13th WSEAS International Conference on COMPUTERS* (2009): 400-05

[9] **Simion G**., Gui V., Otesteanu M. "A Compositional Tehnique for Hand Posture Recognition : New Results." *WSEAS Transaction on Communication* 8 (8) (2009): 805-21

[10] Ommer Björn, Buhmann Joachim "Learning Compositional Categorization Models." *9th European Conference on Computer Vision, Graz, Austria* (2006)

[11] Fei-Fei L., Perona P. "A Bayesian Hierarchical Model for Learning Natural Scene Categories." *Comp. Vis. Patt. Recog.* Ed. IEEEs.

[12] Bruce V., Green P. R., Georgeson M. A. "Visual Perception: Physiology, Psychology, and Ecology." *Psychology Press, East Sussex, UK, 3rd edition* (1996)

[13] Keselman Y., Dickinson S. "Bridging the Representation Gap Between Models and Exemplars." *In IEEE Conf. on Comp. Soc. Work. on Models versus Exemplars in Comp. Vis., Hawaii, USA* (2001)

[14] Binford T. "Visual perception by computer." *IEEE Conference on Systems and Control* (1971)

[15] Agin G., Binford T. "Computer description of curved objects." *IEEE Transactions on Computers* 25 (4) (1976): 439–49

[16] Nevatia R., Binford T. "Description and recognition of curved objects." *Artificial Intelligence* 8 (1977): 77–98

[17] Brooks R. "Model-based 3-D interpretations of 2-D images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 5 (2) (1983): 140–50

[18] Ferrie F., Lagarde J., Whaite P. "Darboux frames, snakes, and superquadrics." *In Proceedings, IEEE Workshop on Interpretation of 3D Scenes* 170–176 (1989)
[19] Gupta A. " Surface and volumetric segmentation of 3D objects using parametric shape models." *Technical Report MS-CIS-91-45, GRASP LAB 128, University of Pennsylvania, Philadelphia* (1991)
[20] Leonardis A., Solina F., Macerl A. "A direct recovery of superquadric models in range images using recover-and-select paradigm." *In Proceedings,Third European Conference on Computer Vision* 800 (1994): 309–18
[21] Pentland A. "Perceptual organization and the representation of natural form." *Artificial Intelligence* 28 (1986): 293–331
[22] Biederman I. " Human image understanding: Recent research and a theory." *Computer Vision, Graphics, and Image Processing* 32 (1985): 29–73
[23] Dickinson S., Pentland A., Rosenfeld A. " A representation for qualitative 3-D object recognition integrating object-centered and viewer-centered models." *In K. Leibovic, editor, Vision: A Convergence of Disciplines. Springer Verlag* (New York, 1990)
[24] Raja N., Jain A. " Recognizing geons from superquadrics fitted to range data." *Image and Vision Computing* 10 (3) (April 1992): 179-90
[25] Dickinson S., Pentland A., Rosenfeld A. "3-D shape recovery using distributed aspect matching." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14 (2) (1992): 174–98
[26] Brooks R. A. " Symbolic reasoning among 3-D models and 2-D images." *Artificial Intelligence* 17 (1981): 285–348
[27] Binford T. O, Levitt T., Mann W. "Bayesian inference in modelbased machine vision. ." *In Proceedings of the Conference on Uncertainty in Artificial Intelligence* (1987)
[28] Huttenlocher D., Ullman S. . "Recognizing solid objects by alignment with an image." *International Journal of Computer Vision* 5 (2) (1990): 195–212
[29] Lowe D. "Perceptual Organization and Visual Recognition." *Kluwer Academic Publishers* (1985)
[30] Forsyth D., Mundy J., Zisserman A., Coelho C., Heller, Rothwell C. "Invariant descriptors for 3d object recognition and pose." *IEEE PAMI* 13: 971-92
[31] Lowe D. G. "The viewpoint consistency constraint." *International Journal of Computer Vision* 1 (1) (1987): 57–72
[32] Ullman S., Basri R. " Recognition by linear combinations of models." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13 (10) (1991): 992–1006
[33] Biederman I. "Recognition-by-components: A theory of human image understanding." *Psychological Review* 94 (2) (1987): 115–47
[34] Riesenhuber M., Poggio T. "Models of object recognition." *Nature Neuroscience* 3 (11) (2000): 1199–204
[35] Turk M., Pentland A. "Eigenfaces for recognition." *Journal of Cognitive Neuroscience* 3 (1) (1991): 71–86,
[36] Murase H., Nayar S. "Visual learning and recognition of 3-D objects from appearance." *International Journal of Computer Vision* (14) (1995): 5-24
[37] Leonardis A., Bischoff H. "Dealing with occlusions in the eigenspace approach." *In Proceedings, IEEE Conference on Computer Vision and Pattern Recognition* (1996): 453–58

[38] Pontil M., Rogai S., VerriA. "Recognizing 3-d objects with linear support vector machines." *In Proceedings of the European Conference on Computer Vision,* (1998): 469–83
[39] Schneiderman H., KanadeT."A statistical method for 3D object detection applied to faces and cars." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2000): 1746–59
[40] Viola P. A., Jones M. J. "Rapid object detection using a boosted cascade of simple features." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2001): 511–18
[41] Jolliffe I. T. "Principle Component Analysis." *Springer -Verlag Berlin Heidelberg* (1986.)
[42] Duda R. O., Hart P. E. , Stork D. G. "Pattern Classification." John Wiley, New York, NY, 2nd edition (2001)
[43] Yuille A.L. , Hallinan P.W. , Cohen.D.S. " Feature extraction from faces using deformable templates." *International Journal of Computer Vision* 8 (2) (1992): 99–111
[44] Jain A. K., Zhong Y, Lakshmanan S. "Object matching using deformable templates." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18 (3) (1996): 267–78
[45] Chui H., Rangarajan A. "A new point matching algorithm for nonrigid registration." *Computer Vision and Image Understanding,* 89 (2-3) (2003): 114–41
[46] Ferrari V., Jurie F., Schmid C. "Accurate object detection with deformable shape models learnt from images." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2007)
[47] Berg A. C., Berg T. L., Malik J. "Shape matching and object recognition using low distortion correspondence." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2005): 26–33
[48] Pentland A., Moghaddam B., Starner T. "View-based and modular eigenspaces for face recognition." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (1994): 84–91
[49] Fisher R. E. "The use of multiple measurements in taxonomic problems." *Annals of Eugenics* 7 (2) (1936): 179–88
[50] Belhumeur P. N., Hespanha J. P., Kriegman D. J. " Eigenfaces versus fisherfaces: Recognition using class specific linear projection." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19 (7) (1997): 711–20
[51] Swain M., Ballard D. " Color indexing " *International Journal of Computer Vision* (1991): 11–32
[52] Schiele B., Crowley J. L. "Recognition without correspondence using multidimensional receptive field histograms." *International Journal of Computer Vision* 36 (1) (2000): 31–52
[53] Vogel J., Schiele B. "Semantic modeling of natural scenes for content-based image retrieval." *International Journal of Computer Vision* 72 (2) (2007): 133–57
[54] Vogel J., Schiele B. "On Performance Characterization and Optimization for Image Retrieval." *Proc. of European Conference on Computer Vision* (2002): 51-55
[55] Harris C., Stephens M. "A combined corner and edge detector." *Alvey Vision Conf*. Ed. Print.

[56] Lowe D. "Object recognition from local scale-invariant features." *Proceedings of the IEEE International Conference on Computer Vision,* (1999): 1150–57
[57] Kadir T., Brady M. "Saliency, Scale and Image Description." *International Journal of Computer Vision* 42 (2) (2001): 83–105
[58] Belongie S., Malik J., Puzicha J. "Shape matching and object recognition using shape contexts." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (24) (2002): 509-22
[59] Berg A. C., Malik J. "Geometric blur for template matching." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2001): 607–14
[60] Hubel D.H., Wiesel T.N. "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex." *J. Physiology* 160 (1962): 106-54
[61] Freeman W., Adelson E. "The design and use of steerable filters." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13 (9) (1991): 891-906
[62] Schaffalitzky F., Zisserman A. "Multi-view matching for unordered image sets." *In Proceedings of the 7th European Conference on Computer Vision* (2002): 414-31
[63] Sivic J., Russell B., Efros A., Zisserman A., Freeman W. "Discovering object categories in image collections." *Proc. Int'l Conf. Computer Vision* (2005)
[64] Csurka Gabriella, Dance Christopher R., Fan Lixin, Willamowski Jutta, Bray Cédric. "Visual Categorization with Bags of Keypoints." *In Workshop on Statistical Learning in Computer Vision, ECCV* (2004): 1-22
[65] Ullman Shimon, Vidal-Naquet Michel, Sal Erez. "Visual features of intermediate complexity and their use in classification." *Nature Neuroscience* 5 (7) (2002): 1-6
[66] Barnard K., Duygulu P., de Freitas N., Forsyth F., Blei D., Jordan M. "Matching words and pictures." *Journal of Machine Learning Research,* (2003): 3:1107–35
[67] Lazebnik S., Schmid C., Ponce J. "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories." *Proc. of IEEE Conference on Computer Vision and Pattern Recognition.* (2006)
[68] Sudderth E., Torralba A., Freeman W., Willsky A. "Learning Hierarchical Models of Scenes, Objects, and Parts." Proc. of International Conference on Computer Vision (2005)
[69] Hofmann T. "Probabilistic latent semantic indexing." *In Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR),* (1999): 50–57
[70] Hofmann T. "Unsupervised learning by probabilistic latent semantic analysis." *Machine Learning,* 42 (1) (2001): 177–96
[71] Dempster A. P., Laird N. M., Rubin D. B. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society* 39 (1977): 1–38
[72] Fergus R., Fei-Fei L., Perona P. and Zisserman A. "Learning Object Categories from Google's Image Search." *Proceedings of the Tenth IEEE International Conference on Computer Vision*. Ed. Print.
[73] Blei D. M. , Ng A. Y. , Jordan M. I. "Latent dirichlet allocation." *Journal of Machine Learning Research,* 3 (2003): 993–1022
[74] Girolami Mark , Kaban Ata "On an Equivalence between PLSI and LDA." *Proceedings of SIGIR* ISBN 1581136463 (2003)

[75] Fischler M. A., Elschlager R. A. "The representation and matching of pictorial structures." *IEEE Transactions on Computers* 22 (1) (1973): 67–92
[76] Lades M., Vorbr¨uggen J. C., Buhmann J. M., Lange J., von der Malsburg C., W¨urtz R. P., Konen W. " Distortion invariant object recognition in the dynamic link architecture." *IEEE Transactions on Computers* 42 (1993): 300–11
[77] Amit Y., Geman D. " A computational model for visual selection." *Neural Computation* 11 (7) (1998): 1691–715
[78] Fukushima K. "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position." *Biological Cybernetics* 36 (4) (1980): 193–202
[79] LeCun Y., Bottou L., Bengio Y., Haffner P. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86 (11) (1998): 2278–324
[80] Riesenhuber M., Poggio T. "Hierarchical models of object recognition in cortex." *Nature Neuroscience* 2 (11) (1999): 1019–25
[81] Serre, T., Riesenhuber, M. "Realistic Modeling of Simple and Complex Cell Tuning in the HMAX Model, and Implications for Invariant Object Recognition in Cortex." *Massachusetts Institute of Technology, Cambridge, MA.* (2004)
[82] Serre T., Wolf L., Poggio T. "Object recognition with features inspired by visual cortex." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2005): 994–1000
[83] Serre T, Wolf L., Bileschi S., Riesenhuber M., Poggio T. "Robust Object Recognition with Cortex-Like Mechanisms." *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE* 29 (3) (2007)
[84] Mutch J., Lowe D. G. "Multiclass object recognition with sparse, localized features." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2006): 11–18
[85] Burl M. C., Weber M., Perona P. "A probabilistic approach to object recognition using local photometry and global geometry." *In Proceedings of the European Conference on Computer Vision* (1998): 628–41
[86] Weber M., Welling M., Perona P. "Towards automatic discovery of object categories." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* (2000): 2101–08
[87] Fergus R., Perona P., Zisserman A. "Object class recognition by unsupervised scale-invariant learning,." *Proc IEEE Conf Computer Vision and Pattern Recognition,* (2003): 264–71.
[88] Burl M.C., Leung T.K., Perona P. "Face localization via shape statistics." *In Int. Workshop on Automatic Face and Gesture Recognition* (1995)
[89] Weber M. "Unsupervised Learning of Models for Object Recognition." *PhD thesis, California Institute of Technology, Pasadena, CA* (2000)
[90] Weber M., Einhauser W., Welling M., Perona P. " Viewpoint-invariant learning and detection of human heads." *In Proc. 4th IEEE Int. Conf. Autom. Face and Gesture Recog., FG2000* (2000): 20-27
[91] Fergus R., Perona P., Zisserman A. "A visual category filter for Google images." *In Proceedings of the European Conference on Computer Vision* (2004): 242–56
[92] Fei-Fei L., Fergus R., Perona P. "One-shot learning of object categories." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28 (4) (2006): 594–611

[93] Agarwal S., Roth D. "Learning a sparse representation for object detection." *In Proceedings of the European Conference on Computer Vision* (2002): 113–30
[94] Agarwal S., Awan A., Roth D. "Learning to detect objects in images via a sparse, part-based representation." *IEEE Trans Pattern Analysis and Machine Intelligence* 26 (11) (2004): 1475–90
[95] Leibe B., Schiele B. "Scale-invariant object categorization using a scale-adaptive mean-shift search." *Pattern Recognition (DAGM), ser. LNCS,* 3175 (2004,): 145–53
[96] Leibe B., Leonardis A., Schiele B. "Combined object categorization and segmentation with an implicit shape model." *In Proceedings of the European Conference on Computer Vision. Workshop on Statistical Learning in Computer Vision* (2004)
[97] Carlson A., Cumby C. , Rosen J., Roth D. . "The snow learning architecture, uiucdcsr-00-2101." *Technical report, Dept. of Computer Science, UIUC* (1999)
[98] Borenstein E., Ullman S. I. " Class-specific, top-down segmentation." *In Proceedings of the 7th European Conference on Computer Vision* (2002): 109-24
[99] Leibe B., Seemann E., Schiele B. "Pedestrian detection in crowded scenes." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2005): 878–85
[100] Seemann E., Leibe B., Schiele B. "Multi-aspect detection of articulated objects." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2006): 1582–88
[101] Felzenszwalb P. F., Huttenlocher D. P. "Efficient matching of pictorial structures." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2000): 66–73
[102] Crandall D. J., Felzenszwalb P. F, Huttenlocher D. P. "Spatial priors for part-based recognition using statistical models." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2005): 10–17
[103] Neagoe Victor, Stanasila Octavian "Recunoasterea formelor si retele neuronale: algoritmi fundamentali " *Matrix Rom* (1999)
[104] Vapnik V. N. "Statistical Learning Theory." *Wiley, New York, NY* (1998)
[105] Zhang H., Berg A. C., Maire M., Malik J. "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2006): 2126–33
[106] Hoiem D., Efros A. A., Hebert M. "Automatic photo pop-up." *ACM Transactions on Graphics* 24 (3) (2005): 577–84
[107] Hoiem D., Efros A. A., Hebert.M. "Geometric context from a single image." *In Proceedings of the IEEE International Conference on Computer Vision and Image Understanding,* (2005): 654–61
[108] Hoiem D., Efros A. A., Hebert M. "Recovering surface layout from an image." *International Journal of Computer Vision* (2007)
[109] Torralba A. "Contextual priming for object detection." *International Journal of Computer Vision* 53 (2) (2003): 169–91
[110] Marr D. "Vision." *W. H. Freeman, San Francisco, CA* (1982)
[111] Bouchard G., Triggs B. " Hierarchical part-based visual object categorization." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2005): 710–15

[112] Epshtein B., Ullman S."Feature hierarchies for object classification." *In Proceedings of the IEEE International Conference on Computer Vision and Image Understanding* (2005): 220–27
[113] Todorovic S., Ahuja N. "Extracting subimages of an unknown category from a set of images." *CVPR* (2006)
[114] Geman S., Potter D., Chi Z. "Composition systems." *Quarterly of Applied Mathematics,* 60 (2002): 707–36
[115] Jin Y., Geman S. "Context and hierarchy in a probabilistic image model." *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (2006)
[116] Keselman Y., Dickinson S. "Generic model abstraction from examples." *CVPR* (2001)
[117] Wang W., Pollak ., Wong T.-S., Bouman C. A., Harper M. P., Siskind J. M. "Hierarchical stochastic image grammars for classification and segmentation." *IEEE Transactions on Image Processing* 15 (10) (2006): 3033–52
[118] Lin L., Peng S. W., Zhu S. C. "An empirical study of object category recognition: Sequential testing with generalized samples." *Proceedings of International Conference on Computer Vision* (2007)
[119] Tu Z. W., Chen X. R., Yuille A. L., Zhu S. C "Image parsing: Unifying segmentation, detection, and recognition." *International Journal of Computer Vision* 63 (2) (2005): 113-40
[120] Olshausen B.A., Field D.J. "Emergence of simple cell receptive field propertyes by learning a sparse code for natural images." *Nature Neuroscience* 381 (1996): 607-09
[121] Blum H. "Biological shapes and visual science." *Jurnal of Theoretical Biology* 38 (1973): 207-85
[122] Tu Z.W., Zu S.C. "Image segmentation for data -driven Markov chain Monte Carlo." *PAMI* (2002)
[123] Werning M., Machery E., Schurz G. "Compositionality of Meaning and Content: Foundational." *Ontos Verlag* 1 (2005)
[124] Frege G. "Meaning and Reference." *Oxford University Press* (1993): 43–45
[125] Katz J. J., Fodor J. A. "The structure of a semantic theory." 39 (1963): 170–210
[126] Ommer Björn. "Learning the Compositional Nature of Objects for Visual Recognition." *Phd Thesis* (2007)
[127] Popa D., **Simion G**., Gui V., Otesteanu M. "Trajectory Based Hand Gesture Recognition." *Computational Intelligence Man-Machine Systems and Cybernetics* (2007)
[128] Popa D., **Simion G**., Gui V., Otesteanu M. "Real Time Trajectory Based Hand Gesture Recognition." *WSEAS Transactions on Information Science & Applications* 5 (4) (2008)
[129] Turk M. "Gesture recognition." *Handbook of Virtual Environments: Design, Implementation, and Applications*. Lawrence Erlbaum Associates, Hillsdale, N.J, 2002. 223-38. Print.
[130] Lenman S., Bretzner L., Thuresson B. "Using marking menus to develop command sets for computer vision based hand gesture interfaces." *NordiCHI '02: Second Nordic Conference on Human- Computer Interaction, ACM Press, New York, NY, USA* (2002): 239–42

[131] Nielsen M., Storring M., Moeslund T.B., Granum E. "A procedure for developing intuitive and ergonomic gesture interfaces for HCI." *In 5th International Gesture Workshop* (2003): 409–20
[132] Lee J., Kunii T. "Constraint-based hand animation." *Models and Techniques in Computer Animation, Springer, Tokyo* (1993): 110–27
[133] Bray M., Koller-Meier E., Gool L.V. "Smart particle filtering for 3D hand tracking." *Sixth IEEE International Conference on Automatic Face and Gesture Recognition* (2004): 675
[134] Bray M., Koller-Meier E., Muller P., Gool L.V., Schraudolph N.N."3D Hand tracking by rapid stochastic gradient descent using a skinning model." *First European Conference on Visual Media Production* (2004): 59–68
[135] Nirei K., Saito H., Mochimaru M., Ozawa S. "Human hand tracking from binocular image sequences." *In 22th International Conference on Industrial Electronics, Control, and Instrumentation* (1996): 297–302
[136] Kuch J.J, Huang T.S "Human computer interaction via the human hand: a hand model." *Twenty-Eighty Asilomar Conference on Signal, Systems, and Computers* (1994): 1252– 56
[137] Rehg J., Kanade T. "Visual tracking of high DoF articulated structures: An application to human hand tracking." *In European Conference on Computer Vision and Image Understanding* (1994): 35–46
[138] Ali Erol, George Bebis, Mircea Nicolescu, Richard D. Boyle, Xander Twombly. "Vision-based hand pose estimation: A review." *Computer Vision and Image Understanding* 108 (2007): 52–73
[139] Heap A. J., Hogg D. C. "Towards 3-D hand tracking using a deformable model." *In 2nd International Face and Gesture Recognition Conference* (1996): 140–45
[140] Wu Y, L. J. Y., Huang T. S. "Capturing natural hand Articulation." *In Proc. 8th Int. Conf. on Computer Vision* 2 (2001): 426–32
[141] Stenger B., Mendonc P. R. S., Cipolla R. "Model-Based 3D Tracking of an Articulated Hand." *Proc. British Machine Vision Conference* 1 (2001): 63-72
[142] http://en.wikipedia.org/wiki/Kalman_filter.
[143] Stenger B. , Thayananthan A. , Torr P.H.S., Cipolla R. "Model-based hand tracking using a hierarchical Bayesian filter." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2006)
[144] Mo Z. , Lewis J.P., Neumann U. "Smartcanvas: a gesture-driven intelligent drawing desk system." *In 10th International Conference on Intelligent User Interfaces, ACM Press* (2005): 239-43
[145] Martin J. , Devin V. , Crowley J.L. "Active hand tracking." *3rd. International Conference on Face & Gesture Recognition, IEEE Computer Society* (1998): 575
[146] Kjeldsen R., Kender J. "Toward the use of gesture in traditional user interfaces." *International Conference on Automatic Face and Gesture Recognition* (1996): 151–56
[147] O'Hagan R.G., Zelinsky A., Rougeaux S. "Visual gesture interfaces for virtual environments." *Interacting with Computers* 14 (2002): 231–50
[148] Segen J., Kumar S. " Gesture VR: Vision-based 3D hand interface for spatial interaction." *Sixth ACM International Conference on Multimedia* (1998): 455–64
[149] Malik S., Laszlo J. "Visual touchpad: a two-handed gestural input device." *6th International Conference on Multimodal Interfaces* (2004): 289–96

[150] von Hardenberg C., Berard F. "Bare-hand human–computer interaction,." *PUI '01: Workshop on Perceptive User Interfaces, ACM Press,* (2001): 1–8
[151] Letessier J., Berard F. "Visual tracking of bare fingers for interactive surfaces." *17th Annual ACM symposium on User Interface Software and Technology* (2004): 119–22
[152] Koike H., Sato Y., Kobayashi Y. "Integrating paper and digital information on enhanceddesk: a method for realtime finger tracking on an augmented desk system." *ACM Transactions on Computer-Human Interaction* 8 (4) (2001): 307–22
[153] Oka K., Sato Y., Koike H. "Real-time tracking of multiple fingertips and gesture recognition for augmented desk interface systems." *Fifth IEEE International Conference on Automatic Face and Gesture Recognition, IEEE Computer Society* (2002)
[154] Pavlovic V. I., Sharma R., HuangT. S. "Visual interpretation of hand gestures for human-computer interaction: A review." *IEEE Trans. on Pattern Recognition and Machine Intelligence,* 19 (7) (1997): 677-95
[155] Freeman W.T., Weissman C.D. "Television Control by Hand Gestures." *Proc. Int'l Workshop on Automatic Face and Gesture Recognition, Zurich, Switzerland* (1995): 179-83
[156] Quek F.K.H. "Eyes in the Interface." *Image and Vision Computing* 13 (1995)
[157] Kolsch M., Turk M. "Robust hand detection." *Sixth IEEE International Conference on Automatic Face and Gesture Recognition* (2004)
[158] Ong E.-J., Bowden R. "A boosted classifier tree for hand shape detection." *Sixth IEEE International Conference on Automatic Face and Gesture* (2004): 889–94
[159] Wu Y., Huang T. "View-independent recognition of hand postures." *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2 (2000): 88–94
[160] Sato Y., Saito M., Koik H. "Real-time input of 3D pose and gestures of a user's hand and its applications for HCI." *IEEE Virtual Reality Conference, IEEE Computer Society,* (2001)
[161] Jo K.H., Kuno Y., Shirai Y. "Manipulative hand gesture recognition using task knowledge for human computer interaction." *3rd. International Conference on Face & Gesture Recognition,* (1998)
[162] Abe K., Saito H., Ozawa S. . "3D drawing system via hand motion recognition from two cameras." *IEEE International Conference on Systems, Man, and Cybernetics* 2 (2000): 840–45
[163] Utsumi A., Ohya J. "Multiple-hand-gesture tracking using multiple cameras." *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (1999): 473–78
[164] O'Hagan R., Zelinsky A. "Finger track—a robust and real-time gesture interface." *10th Australian Joint Conference on Artificial Intelligence* (1997): 475–84
[165] Crowley J., Berard F., Coutaz J. " Finger tracking as an input device for augmented reality." *International Workshop on Gesture and Face Recognition* (1995): 195–200
[166] Pavlovic V.I., Sharma R., Huang T.S. "Invited speech: gestural interface to a visual computing environment for molecular biologists." *2nd International Conference on Automatic Face and Gesture Recognition, IEEE Computer Society* (1996)

[167] Martin J., Devin V., Crowley J.L. "Active hand tracking." *3rd. International Conference on Face & Gesture Recognition, IEEE Computer Society* (1998)
[168] Maggioni C., Gesturecomputer—history, design and applications." *in: R. Cipolla, A. Pentland (Eds.), Computer Vision for Human-Machine Interaction, Cambridge* (1998): 312–25
[169] Jennings C. "Robust finger tracking with multiple cameras." *International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems* (1999)
[170] Davis J., Shah M. "Toward 3D gesture recognition." *International Journal of Pattern Recognition and Artificial Intelligence* 13 (3) (1999): 381–93
[171] Kim H., Fellner D.W. "Interaction with hand gesture for a backprojection wall." *Computer Graphics International, IEEE Computer Society* (2004): 395–402
[172] Roberts L. G. "Machine perception of three-dimensional solids." *Optical and Electro-Optical Information Processing* (1965): 159-97
[173] Pingle K. K. "Visual perception by computer." *Automatic Interpretation and Classication of Images* (1969): 277-84
[174] Prewitt J. M. S. "Object enhancement and extraction." *Picture Processing and Psychophysics* 75-149 (1970)
[175] Haralick R. M. "Digital step edges from zero-crossings of second directional derivatives." *IEEE Trans. Pattern Analysis and Machine Intell* 6 (1984)
[176] Canny J. "Finding edges and lines in images." *Tehnical Reoprt AITR-720, Massachusetts Institute of Technologie* (1983)
[177] Canny J. "A computational approach to edge detection." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8 (6) (1986): 679–98
[178] Tagare H. D., deFigueiredo R. J. P. "On the localization performance measure and optimal edge detection." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12 (12) (1990): 1186–90
[179] Schmid C., Mohr R., Bauckhage C. "Evaluation of interest point detectors." *International Journal of Computer Vision* 37 (2) (Feb. 2000): 151–72
[180] Smith S.M., Brady J.M. "SUSAN—A new approach to low level image processing." *International Journal of Computer Vision* 23 (1) (1997): 45–78
[181] Lowe, D. G. " Distinctive image features from scale-invariant keypoints." *International Journal of Computer Vision* 60 (2) (2004): 91–110
[182] Attneave F. "Some informational aspects of visual perception." *Psychological Review* 61 (3) (1954): 183–93
[183] Goldstein E. B. "Sensation and Perception." *Wadsworth, Belmont, CA, 3rd edition* (1989)
[184] M., Wertheimer. "A Source Book of Gestalt Psychology, New York, NY: Harcourt, Brace." *English translation in W.D. Ellis, editor* (1938)
[185] Viola P. A., Jones M. J. "Robust real-time face detection." *Robust real-time face detection. International Journal of Computer Vision* 52 (7) (2004): 137–54
[186] Fei-Fei L., Fergus R., Perona P. "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories,." *In Proc IEEE Conf Computer Vision and Pattern Recognition. Workshop on Generative Model Based Vision* (2004)
[187] Partzen Emanuel. "On estimation of a probability density function and mode " *Annals of Mathematical Statistics* 33 (3) (1962): 1065-76

[188] Hampel F. R. , Rousseeuw P. J. , Ronchetti E., Stahel W. A. "Robust Statistics: The Approach Based on Infuence Functions." *John Wiley, New York* (1986)
[189] Huber P. J. . "Robust Statistics." *John Wiley, New York* (1981)
[190] Lindgren B. W. "Statistical Theory." *Chapman and Hall, London* (1993)
[191] Comaniciu D., Meer P. "Mean Shift : A robust approach toward feature space analysis." *IEEE Trans. on Pattern Recognition and Machine Intelligence* 24 (5) (2002)
[192] Anil K. Jain, Robert P.W. Duin, Jianchang Mao. "Statistical Pattern Recognition: A Review." *Ieee Transactions On Pattern Analysis And Machine Intelligence* 22 (1) (2000)
[193] Schiele, B. "PhD thesis, Object Recognition Using Multidimensional Receptive Field Histograms.", 1997
[194] Asada H., Brady, M. "The curvature primal sketch." *EEE Transactions on Pattern Analysis and Machine Intelligence* 8 (1) (1986): 2–14.
[195] Mokhtarian F., Suomela R. "Robust image corner detection through curvature scale space." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (12) (1998): 1376-81
[196] Tomasi C., Kanade T. "Detection and tracking of point features." *Technical Report, Carnegie Mellon University* CMU-CS-91-132 (1991)
[197] T. Kadir, M. Brady. "Saliency, Scale and Image Description." *International Journal of Computer Vision* 42 (2) (2001): 83–105
[198] Ke Y., Sukthankar R. "PCA-SIFT: A more distinctive representation for local image descriptors." *Proceedings of the Conference on Computer Vision and Pattern Recognition* (2004): 511-17
[199] Mikolajczyk K., Schmid C. "A Performance Evaluation of Local Descriptors." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27 (10) (2005)
[200] Jones J.P., Palmer L.A. "An Evaluation of the Two-Dimensional Gabor Filter Model of Simple Receptive Fields in Cat Striate Cortex." *J. Neurophysiology* 58 (1987): 1233-58
[201] Wyszecki G., Stiles W. S. "Color Science: Concepts and Methods, Quantitative Data and Formulae (" (1982)
[202] Skarbek W., Koschan A. "Colour image segmentation – a survey." Tech. rep., Institute for Technical Informatics, Technical University of Berlin, (1994)
[203] Gui V., Alexa F., Caleanu C., Fuiorea D. "Motion segmentation and analysis in video segmentation." *WSEAS Transaction on Circuits and System* 6 (1) (2007): 142-48
[204] Judd D., Mckinley P., Jain A.K."Large-Scale Parallel Data Clustering,." *IEEE Trans. Pattern Analysis and Machine Intelligence* 20 (8) (1998): 871-76
[205] Bhatia S.K., Deogun J.S. "Conceptual Clustering in Information Retrieval." *IEEE Trans. Systems, Man, and Cybernetics* 28 (3) (1998): 427-36
[206] Frigui H., Krishnapuram R. "A Robust Competitive Clustering Algorithm with Applications in Computer Vision." *IEEE Trans. Pattern Analysis and Machine Intelligence* 21 (5) (1999): 450-65
[207] Abbas H.M., Fahmy M.M. "Neural Networks for Maximum Likelihood Clustering." *Signal Processing* 36 (1) (1994): 111-26
[208] Carpineto C., Romano G. "A Lattice Conceptual Clustering System and Its Application to Browsing Retrieval." *Machine Learning* 24 (2) (1996): 95-122

[209] Ommer Björn, Buhmann Joachim. "Learning the Compositional Nature of Visual Object Categories for Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence* IEEE computer Society Digital Library. IEEE Computer Society (2009)
[210] Winkler G. "Image Analysis, Random Fields and Dynamic Monte Carlo Method A Mathematical Introduction." *Springer -Verlag Berlin Heidelberg* ISBN 3-540-57069-1 (1995)
[211] Wang C C , Wang K C. "Hand Posture recognition using Adaboost with SIFT for human robot interaction." 370 (ISSN 0170-8643) (2008)
[212] Chen Qing , Georganas N.D., Petriu E.M. "Real-time Vision based Hand Gesture Recognition Using Haar-like features." *IEEETransactions on Instrumentation and Measurement* (2007)
[213] Just A., Rodriguez Y., Marcel S. "Hand Posture Classification and Recognition using the Modified Census Transform." *Automatic Face and Gesture Recognition* (2006)