A Compositional Technique for Hand Posture Recognition: New Results

GEORGIANA SIMION, VASILE GUI, MARIUS OTESTEANU
Department of Communication
"Politehica" University of Timisoara
Bd. V. Parvan, Nr. 2,300223 Timisoara
ROMANIA
georgiana.simion@etc.upt.ro, vasile.gui@etc.upt.ro, marius.otesteanu@etc.upt.ro

Abstract: New results obtain with the compositional technique for hand posture recognition is presented. Compositional methods are a powerful approach in image understanding. Most papers using this concept address image categorization problems. We recently propose a hand pose recognition method using the compositional approach. In this paper we present further development of our method and new results.

Key-Words: - compositional technique, hand posture recognition, sparse, clustering

1 Introduction

Humans perform various gestures in their daily life. Hand gestures are a powerful communication modality, they are natural and intuitive.

Gesture recognition is nowadays an active topic of vision research which has applications in diverse fields such as: surveillance, sign language translation, interactive games, performance analysis, monitoring, and remote control of home appliances, virtual reality, disability support, medical systems, and many others.

There has been a great emphasis lately in HCI research to create easier to use interfaces by directly employing natural communication and manipulation skills of humans. Among different body parts, the hand is the most effective, general-purpose interaction tool due to its dexterous functionality in communication and manipulation. Various interaction styles tend to import both modalities to allow intuitive, natural interaction.

There are different approaches for hand gesture analysis. Gestures can be classified as static and dynamic gestures. A static gesture is a particular hand configuration and pose represented by a single image. A dynamic gesture is represented by a sequence of images. A solution to capture the richness of a hand gesture is to use Dataglove devices. These devices are able to capture the fingers position and movement of the hand but require these special gloves, which are not accepted as a solution for many applications.

Vision based hand gestures recognition techniques can be divided in: 3D model based and appearance based approaches. The 3D hand model based approach uses the 3D kinematics' hand model with a degree of freedom and try to estimate the hand parameters by comparison between the input images and the possible 2D appearance projected by the 3-D and model [1], [2], [3], [4]. A simple appearance based approach is to look for skin colored regions in an image to segment the hand. The HSV space is preferred but obviouslyproblems show up when skin like objects there are in scene. Zhou [5] presented a bottom-up algorithm for posture recognition, based on local orientation histograms features; the advantage is a higher recognition accuracy but the local orientation histograms features are affected by rotation. The recent vision literature increased interest in approaches working with local invariant features [6], [7], [8], [9].

According to [10] different approaches to hand pose estimation can be seen in figure1.

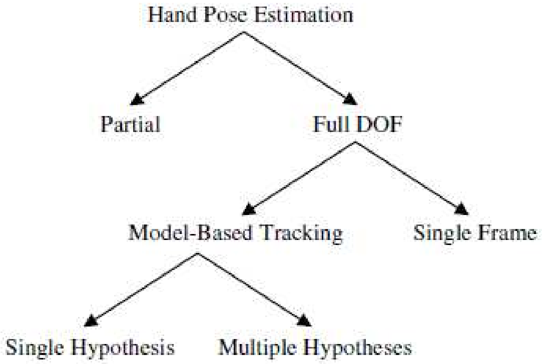


FIG.1 DIFFERENT APPROACHES TO HAND POSE ESTIMATION

By single frame pose estimation they mean estimating the pose of the hand using a single image or multiple images taken simultaneously from different views. One motivation for addressing this more challenging problem is for the purpose of initializing tracking without imposing too many constraints on the user.

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[12] learned object categories from Google's image search.

Generally, object recognition approaces consist of four major stages: (1) feature detection, (2) object representation, (3) training, and (4) recognition. This work 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 representatively features for sparse representation.

For the second stage – object representation – most approaches partition extracted features into clusters, also known as parts. This 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.

The main differences between the classical model for statistical pattern recognition and compositional model consist of the intermediate layer of abstraction introduced by the compositional ones.

2 Theoretical Backgrounds

The Principle of Compositionality and modeling decisions

According to [13] 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 [14]. *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 [15] . 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 does 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 [16] 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.

In literature there are presented different visual recognition methods based on different levels of semantic granularity. The granularity defines the complexity of the recognition task and the interclass variability. The current approaches to object categorization can be characterized according to the modeling decisions they take.

Local Descriptors: There are many methods to capture image region information: appearance patches [17], SIFT features [18], geometric blur,Gabor filters, localized histograms [19], [20], and edge contours based methods .

Spatial Model The bag of features methods offer robustness with respect to alteration of individual parts of an object at low computational costs, but it fails to capture any spatial relations between local image patches and usually adapt to background features. On the other end there are constellation models.

Hierarchies The research on object recognition has aimed to build hierarchical models, despite this, many popular methods such as [17] are single layered. Recently, probabilistic latent semantic analysis (pLSA) [21] and latent dirichlet allocation [22] are used to introduce a hidden representation layer of abstract concepts [23], [12].

Learning Paradigm: Another modeling decision is related to the learning paradigm, although discriminative approaches have been shown to yield superior performance in the limited case of large training sets, generative models have been very popular in the vision community.

3 Problem Solution

Finding good sparse features

Even if the proposed method is a general one, for different applications it is still important what features are used [24] for the sparse hand gesture representation. The potential benefits of feature selection include, first and foremost, better accuracy of the inference engine and improved scalability (defining the curse of dimensionality). Secondary benefits include better data visualization and understanding, reduce measurement and storage requirements, and reduce training and inference time.

The ideal features are not affected by occlusion and clutter, there are invariant (or covariant), there are also robust, it means that the 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 quantity point of view many features can be generated for even small objects and offer a precise localization.

The first question according to compositionality technique 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 finger tips 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?

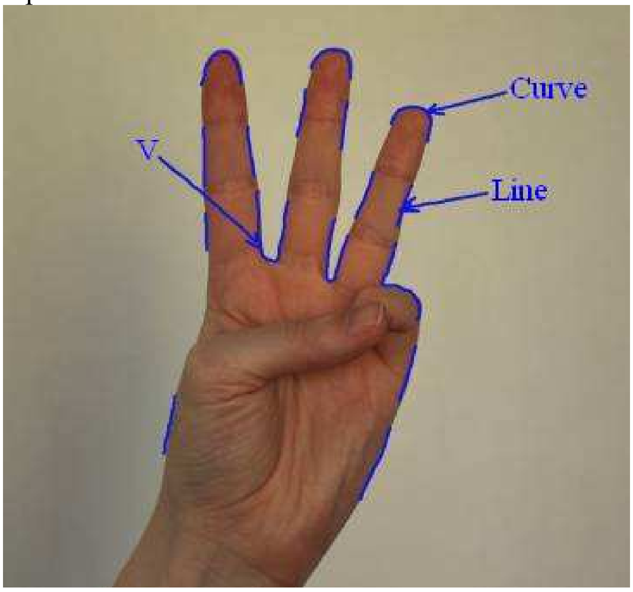
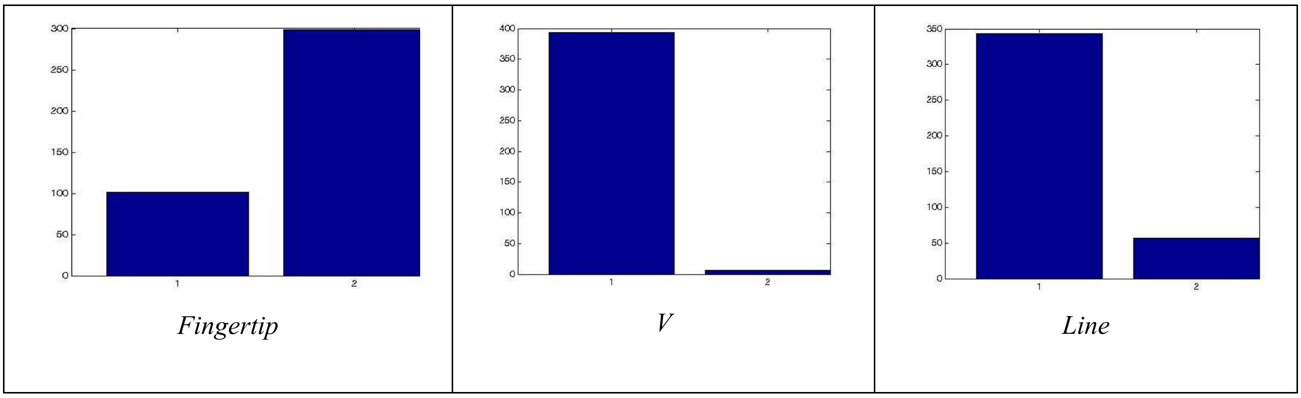


FIG.2 V SHAPE, CURVE AND LINE

In this work in order to capture the salient regions Harris corner detector was used on hand contours. Comparative tests carried out amongst different approaches demonstrated that detectors which combine the smoothed image derivates via the autocorrelation matrix are robust and achieve some of the best results [25].

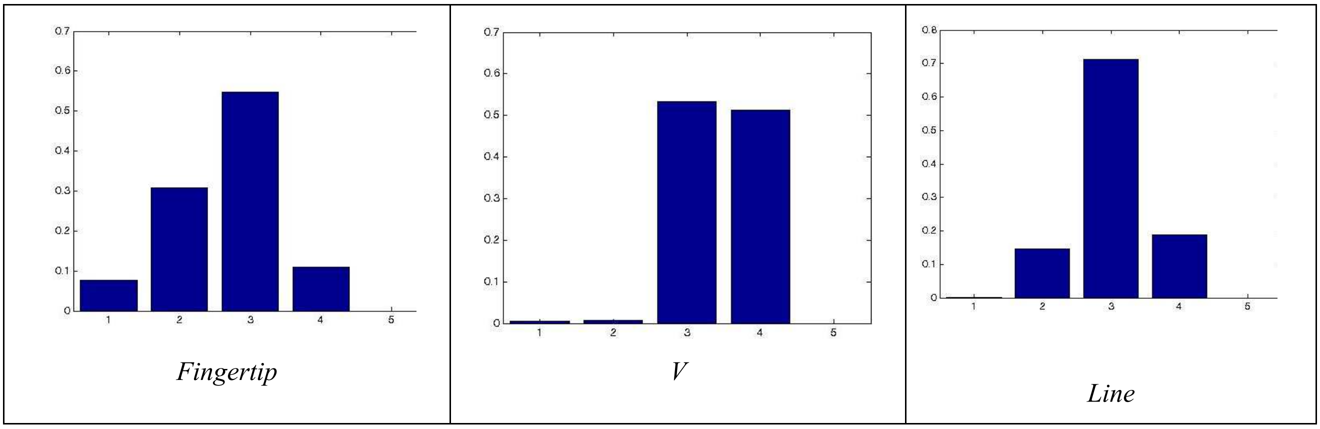
Perhaps the most prominent candidate in this category is the edge and corner detector proposed by Harris and Stephens [26]. The Harris corner detector is based on the local auto-correlation function of a signal. The local auto-correlation function measures the local changes of the signal with patches shifted by a small amount in different directions. 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.



Fingertip *V* *Line*

FIG. 3 COLOR HISTOGRAM FOR PATCHES SELECTED AROUND INTEREST POINTS DETECTED IN DIFFERENT AREAS OF HAND



Fingertip *V* *Line*

FIG. 4 ORIENTATION HISTOGRAM FOR PATCHES SELECTED AROUND INTEREST POINTS DETECTED IN DIFFERENT AREAS OF HAND

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 [27], [28] should not disturb, so to avoid as much as possible its influence, color histogram with two bins (skin – non skin) are used.

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. 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.

Quadratic patches of size 20 × 20 pixels are extracted around each Harris interest point to capture discriminative local information. 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, ones of the most popular color spaces are: RGB, HIS, HSV, HSL, TLS, YCrCb. 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 [29]: 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.

The resulting seven parameters extracted from a patch are used to form a feature vector, $\mathbf{e}_i$. It is important to remark the small dimension of the feature vector, which is seven. An example of interest points detected with Harris interest point detector on the hand contour, extracted with Canny edge detector can be seen in Fig.5

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.

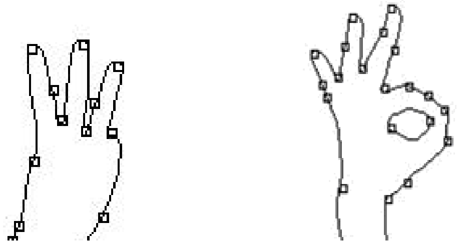


FIG.5 EXAMPLES OF HARRIS INTEREST POINTS DETECTED ON HANDS CONTOURS

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 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.

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 [30], information retrieval [31], image segmentation [32], signal compression and coding [33] , and machine learning [34]. 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 [35], 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 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 [34] 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 [20]. The probability measures from Eq. 1 are always positive and hence random fields.

According to the Gibbs distribution low, the feature assignment random variable, F_i, is given by Eq. 1

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

$$Z(\mathbf{e}_i) = \sum_{\nu} \exp(-d_{\nu,\sigma}(\mathbf{e}_i)) \qquad (2)$$

$$d_{\nu,\sigma}(\mathbf{e}_i) = \left\| \mathbf{e}_i - \mathbf{a}_\nu \right\|^2 \qquad (3)$$

where F_i is a feature assignment random variable, $P(F_i = \nu \mid \mathbf{e}_i)$ is the probability of feature vector $\mathbf{e}_i$ to belong to the class ν defined by the prototype vector $\mathbf{a}_\nu$, $d_\nu(\mathbf{e}_i)$ is the Euclidian distance of a measured feature $\mathbf{e}_i$ to a centroid $\mathbf{a}_\nu$ of class and is a normalization factor. Eq.1 is evaluated for all centroids $\mathbf{a}_\nu$ and the results for a feature point described by ei are grouped in a part distribution vector:

$$\mathbf{d}_i = \left(P(F_i = 1 \mid \mathbf{e}_i), \ldots P(F_i = k \mid \mathbf{e}_i) \right)^T \qquad (4)$$

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. In this work the number of pixels N is 25. This number depends on the types of objects and compositions that one's want 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 [19] 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.1. If $\Gamma_j = \{\mathbf{e}_1,...\mathbf{e}_m\}$ denotes the grouping of parts represented by $\mathbf{e}_1,...,\mathbf{e}_m$, and $\mathbf{d}_1,...,\mathbf{d}_m$, the compositions are then represented by the vector valued random variable G_j which is a bag of parts with the particular values given by:

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

(5)

where the number of constituents, $m=\left|\Gamma_j\right|$, 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 **gj**, 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.

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, gj, are clustered using once more k-means clustering. Let $\boldsymbol{\pi}_i \in \Pi$ be the composition prototypes found by clustering. Then the prior assignment probabilities of candidate compositions to clusters $P(\boldsymbol{\pi}_i)$, are computed using the Gibbs distribution:

$$P(\pi_i = \Pi \mid \mathbf{g}_j) = Z(\mathbf{g}_j)^{-1}\exp(-d_{\Pi,\sigma}(\mathbf{g}_j)) \qquad (6)$$

$$Z(\mathbf{g}_j) = \sum_{\Pi}\exp(-d_{\Pi,\sigma}(\mathbf{g}_j)) \qquad (7)$$

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:

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

$$P(c \mid \Gamma_j) \approx \frac{P(\Gamma_j \mid c)}{\sum_c P(\Gamma_j \mid c)} \qquad (8)$$

where $c \in \wp$, $\wp$ is the set of all category labels. The category posterior is used to calculate the relevance of a composition for discriminating categories. 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 categories is then estimated by the entropy of its category posterior:

$$H(P_{\Gamma_j}) = -\sum_{c\in\wp} P(c \mid \Gamma_j)\log P(c \mid \Gamma_j) \qquad (9)$$

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 categories 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 [19] Ommer and Buhmann proposed the following cost function:

$$S(\boldsymbol{\pi}_i) \propto -P(\boldsymbol{\pi}_i) + \lambda H(P_{\boldsymbol{\pi}_i}) \qquad (10)$$

In [36] Ommer proposed the cost function described by Eq. 11

$$S(\boldsymbol{\pi}_i) \propto -\log P(\boldsymbol{\pi}_i) + \lambda H(P_{\boldsymbol{\pi}_i}) \qquad (11)$$

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 [36] it is proposed a method to compute the value of the parameter λ. The estimation of parameter proposed by Ommer is not a robust one because it uses the max and min values which are sensitive to outliers, as one can see in equation

$$\lambda = \frac{\max_i \log P(\boldsymbol{\pi}_i) - \min_i \log P(\boldsymbol{\pi}_i)}{\max_i H(P_{\boldsymbol{\pi}_i}) - \min_i H(P_{\boldsymbol{\pi}_i})} \qquad (12)$$

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

$$\lambda = \frac{IQR(\log P(\boldsymbol{\pi}_i))}{IQR(H(P_{\boldsymbol{\pi}_i}))} \qquad (13)$$

From the set of all compositional prototypes a set of relevant composition prototypes is established through minimization of Eq.11. For all composition prototypes 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 computation is better than one obtained using Ommer equation; these results are presented in the experimental part.

Training step

The training procedure is carried out according to the diagram shown in Fig.7. For all training images the features vectors 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 and the resulted composition prototypes are used to form the composition codebook. Based on the cost function defined in Eq. 11 relevant compositions prototypes are learned in the next stage. 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 . In figure below an example of relevant composition detected for class f are shown.

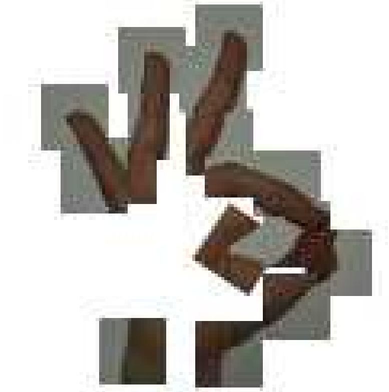


FIG. 6 EXAMPLE OF RELEVANT COMPOSITIONS

Hand posture Recognition

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.

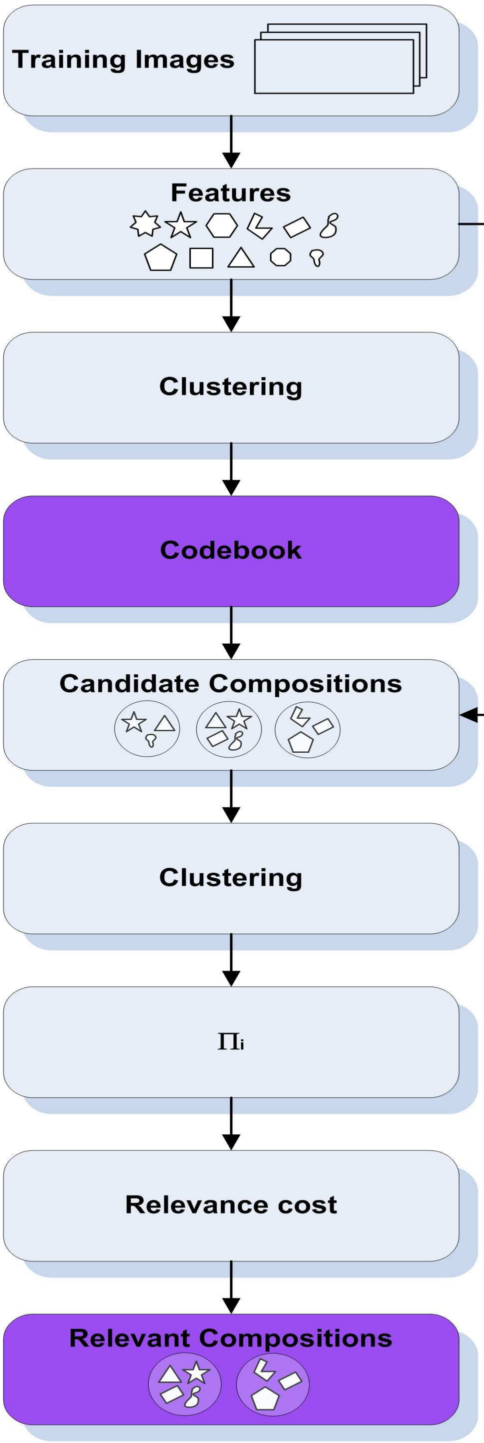


FIG. 7 TRAINING DIAGRAM

$$\mathbf{h}_i = \begin{bmatrix} x_i \\ y_i \\ \mathbf{g}_j \end{bmatrix} = \begin{bmatrix} \alpha x_r \\ \alpha y_r \\ \left(P(F_i = 1 \mid \mathbf{e}_i), \ldots P(F_i = k \mid \mathbf{e}_i)\right)^T \end{bmatrix} \quad (14)$$

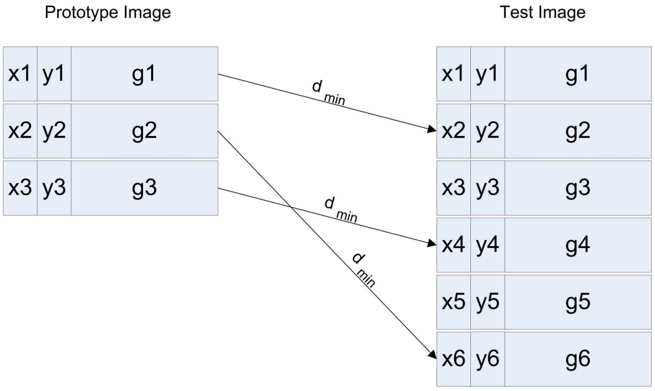
Where $x_i = \alpha(x - x_{median}) = \alpha x_r$

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

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.

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 neuro-networks. For each new image only the minimum distance from the prototypes image compositions to test image composition $\mathbf{h}_i$ is computed $\min\limits_{c_i} \left\| h_\nu^{k,q_\nu} - h_i^{c_i} \right\|$, then all these distances are sum and normlized.



x1	y1	g1
x2	y2	g2
x3	y3	g3

x1	y1	g1
x2	y2	g2
x3	y3	g3
x4	y4	g4
x5	y5	g5
x6	y6	g6

FIG.8 MINIMUM DISTANCE BETWEEN COMPOSITIONS FOUND IN PROTOTYPE IMAGE AND THOSE FOUND IN THE TEST IMAGE

ν is the number of picture per class, k is the class, q_ν is the number of compositions from a class, i is the curent image and c_i is the number of composition for the test image.

The reason why the distance from the test image compositions to prototypes image 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 prototype images.

These distances are computed for all prototypes images.

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

$$\arg\min_{k_v}\left\{\frac{1}{\#q_v}\sum_{q_v}\min_{c_i}\left\|\mathbf{h}_v^{k,q_v}-\mathbf{h}_i^c\right\|\right\}=-d(c,k) \quad (15)$$

4 Experiments

In order to evaluate the compositional approach to hand gesture recognition a set of nine hand postures is used. The set of hand posture represents nine hand postures from ASL((American Sign Language) more exactly letters: a, c, d, e, f, p, u, w, x. Pictures were taken using a Nikon D60 camera.

We have 30 training images per class and a white wall as background. Pictures were taking in natural conditions, no artificial light was added. The pictures were taken in two different days; this can be notice from illumination changes, which is not the same for all pictures.

The images resolution is 255 x 171. 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 varies in experiments.

The two bin colour histogram is extract only for the red component of the RGB image. Parameter σ from Eq.1 is equal to 1 and parameter σ from Eq.2 equal to 0.05.

The number of composition prototypes 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 composition result from interest points detected on hand and just a few are the result of some points detected on background.

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

Experiment 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 determinates farther the relevant composition used to describe an image.

Based on the proposed equation in [20], which is sensitive to noise an using the cost function,

$$S(\boldsymbol{\pi}_i)\propto-\log P(\boldsymbol{\pi}_i)+\lambda H(P_{\pi_i}) \quad (16)$$

$$\lambda=\frac{\max_i \log P(\pi_i)-\min_i \log P(\pi_i)}{\max_i H(P_{\pi_i})-\min_i H(P_{\pi_i})} \quad (17)$$

the following results are obtained: the error rate is 7.037% and confusion matrix can be seen in Fig.9. The diagonal is 93.033%

Class / Predeicted	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**

FIG. 9 THE CONFUSION MATRIX FOR SET 1 WITH 19 RELEVANT COMPOSITION , $\alpha = 0.02$, λ COMPUTED USING OMMER EQUATION.

Using the proposed robust estimation of the parameter λ an the cost function:

$$S(\boldsymbol{\pi}_i)\propto-\log P(\boldsymbol{\pi}_i)+\lambda H(P_{\pi_i})$$

$$\lambda=\frac{\mathrm{IQR}(\log P(\boldsymbol{\pi}_i))}{\mathrm{IQR}(H(P_{\boldsymbol{\pi}_i}))}$$

The following results are obtained: the error rate is 3.70% and confusion matrix can be seen in Fig.10. The diagonal is 96.29%.

Class Predeicted	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.

FIG. 10THE CONFUSION MATRIX FOR SET 1 WITH 19 RELEVANT COMPOSITION PROTOTYPES, $\alpha = 0.02$, λ COMPUTED USING THE PROPOSED EQUATION

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

Experiment 2: The importance of parameter α

As it was mention above the 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 experiment 1 the value of parameter α is changed to 0.1. The results are summarized in the next confusion matrix. The diagonal is 93.4%, the error rate is 6.6%, and only one hand posture is recognized 100%.

Class Predeicted	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

FIG.11THE CONFUSION MATRIX FOR SET 1 WITH 19 RELEVANT COMPOSITION PROTOTYPES, $\alpha = 0.1$, λ COMPUTED USING THE PROPOSED EQUATION

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

Class Predeicted	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

FIG. 12 THE CONFUSION MATRIX FOR SET 1 WITH 19 RELEVANT COMPOSITION PROTOTYPES, $\alpha = 0.5$, λ COMPUTED USING THE PROPOSED EQUATION

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

These experiments prove the importance of parameter α, its best value is learned from the training data.

Experiment 3: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%

Class Predeicted	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

FIG. 13 THE CONFUSION MATRIX FOR SET 1 WITH 14 RELEVANT COMPOSITION PROTOTYPES, $\alpha = 0.02$, λ COMPUTED USING THE PROPOSED EQUATION, K-MEANS CLUSTERING ALGORITHM

For 16 relevant compositon prototypes, keeping all the other parameters at the same value $\alpha = 0.02$, k-means clustering, the error rate is 7%, the diagonal is 93%.

Class Predeicted	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

FIG.14 THE CONFUSION MATRIX FOR SET 1 WITH 16 RELEVANT COMPOSITION PROTOTYPES, $\alpha = 0.02$, λ COMPUTED USING THE PROPOSED EQUATION, K-MEANS CLUSTERING ALGORITHM.

For 18 relevant compositon prototypes, keeping all the other parameters at the same value $\alpha = 0.02$, k-means clustering, the error rate is 4.4%, the diagonal is 95.6 %.

Class Predeicted	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

FIG.15 THE CONFUSION MATRIX FOR SET 1 WITH 18 RELEVANT COMPOSITION PROTOTYPES, $\alpha = 0.02$, λ COMPUTED USING THE PROPOSED EQUATION, K-MEANS CLUSTERING ALGORITHM

In Fig.16 and Fig.5.17 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.

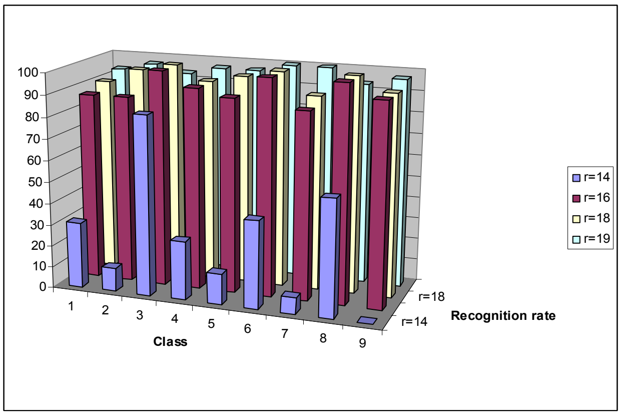


FIG.16 RECOGNITION RATE PER CLASS FOR DIFFRENT NUMBERS OF RELEVANT COMPOSITION PROTOTYPES

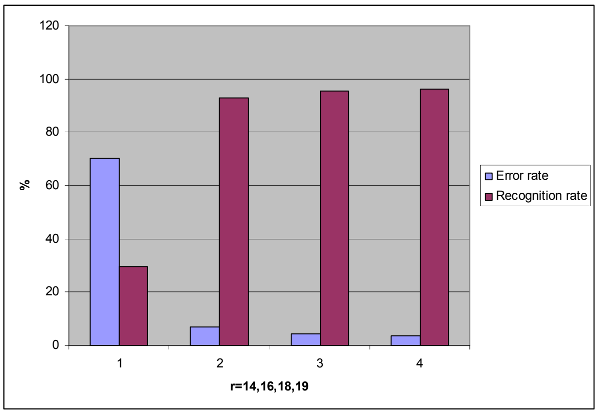


FIG.5.17 THE EVOLUTION OF ERROR RATE AND RECOGNITION RATE FOR R=16,18 AND 19

5 Conclusions

In this work 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 [20], this work designed the processing modules by considering the specifics of the hand gesture recognition problem, where needed.

The results obtained for hand pose classification are better than results reported in object categorization using compositional approaches. Fei-Fei [20] used this technique to recognize 13 natural scene categories; the average performance reported was 64%. Fergus [21] classified 7 categories: airplane, cars rear, face, guitar, leopard, motorbike; the average performance reported was 72%. It is difficult to compare our results with the results obtained in image categorization due to the fact that these classes widely differs, while our classes consist of hand gestures.

In the same time our results are similar to the best results reported in hand gesture recognition with alternative approaches. Since in the compositional approach there are many optimizations left unexplored by our work, we consider the current results as promising and the compositional approach in hand gesture recognition a subject deserving further research work. In our future work we will look for new sparse features that help to discriminate better between hand postures.

References

[1] K. T. Rehg J.M., "Visual tracking of high DOF articulated structures: An application to human hand tracking," *Proc. European Conference on Computer Vision,* 1994.

[2] H. D. C. Heap A. J., "Towards 3-D hand tracking using a deformable model," *In 2nd International Face and Gesture Recognition Conference,* pp. 140–145, 1996.

[3] M. P. R. S. Stenger B. , Cipolla R. , "Model-Based 3D Tracking of an Articulated Hand," *Proc. British Machine Vision Conference,* vol. 1, pp. 63-72, 2001.

[4] T. A. Stenger B. , Torr P.H.S., Cipolla R., "Model-based hand tracking using a hierarchical Bayesian filter," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2006.

[5] L. D. J. Zhou Hanning, Huang Thomas S., "Static Hand Gesture Recognition based on Local Orientation Histogram Feature Distribution Model," 2004.

[6] W. K. C. Wang C C "Hand Posture recognition using Adaboost with SIFT for human robot interaction," vol. 370, 2008.

[7] M. J. Lienhart R. , "An extended set of Haar-like features for rapid object detection," *Proc. IEEE Int. Conf. Image Process,* vol. 1, pp. 900–903, 2002.

[8] B. Andre L. C., Farhad Dadgostar, "Real-time hand tracking using a set of cooperative classifiers based on Haar-like features," *Res. Lett. Inf. Math. Sci.,* vol. 7, pp. 29-42, 2005.

[9] G. N. D. Chen Qing , Petriu E.M, "Real-time Vision based Hand Gesture Recognition Using Haar-like features," *IEEETransactions on Instrumentation and Measurement,* 2007.

[10] G. B. Ali Erol , Mircea Nicolescu, Richard D. Boyle, Xander Twombly, "Vision-based hand pose estimation: A review," *Computer Vision and Image Understanding* vol. 108, pp. 52–73, 2007.

[11] P. P. Fei-Fei L., "A Bayesian Hierarchical Model for Learning Natural Scene Categories," in *Comp. Vis. Patt. Recog.*
, IEEE, Ed., 2005.

[12] F.-F. L. Fergus R., Perona P., Zisserman A., "Learning Object Categories from Google's Image Search," in *Proceedings of the Tenth IEEE International Conference on Computer Vision*. vol. 2, 2005, pp. 1816 – 1823.

[13] Attneave F., "Some informational aspects of visual perception," *Psychological Review,* vol. 61, pp. 183–193, 1954.

[14] Biederman I., "Recognition-by-components: A theory of human image understanding," *Psychological Review,* vol. 94, pp. 115–147, 1987.

[15] M. E. Werning M. , Schurz G. , "Compositionality of Meaning and Content: Foundational," *Ontos Verlag,* vol. 1, 2005.

[16] Goldstein E. B., "Sensation and Perception," *Wadsworth, Belmont, CA, 3rd edition,* 1989.

[17] P. P. Fergus R. , Zisserman A., "Object class recognition by unsupervised scale-invariant learning,," *Proc IEEE Conf Computer Vision and Pattern Recognition,,* pp. 264–271., 2003.

[18] D. G. Lowe, " Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision,* vol. 60, pp. 91–110, 2004.

[19] B. J. Ommer Björn, "Learning Compositional Categorization Models," *9th European Conference on Computer Vision, Graz, Austria* 2006.

[20] B. J. Ommer Björn, "Learning the Compositional Nature of Visual Object Categories for Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. IEEE computer Society Digital Library. IEEE Computer Society, 2009.

[21] Hofmann T., "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning,,* vol. 42, pp. 177–196, 2001.

[22] N. A. Y. Blei D. M. , Jordan M. I., "Latent dirichlet allocation," *Journal of Machine Learning Research,,* vol. 3, pp. 993–1022, 2003.

[23] R. B. Sivic J. , Efros,A. , Zisserman A.,

Freeman, W., "Discovering object categories in image collections.," *Proc. Int'l Conf. Computer Vision,* 2005.

[24] G. Caleanu C., Alexa F, "Direct Search Optimized Feature Extraction," *WSEAS Transaction on System and Control,* vol. 1, pp. 113-120, 2006.

[25] M. R. Schmid C., Bauckhage C., "Evaluation of interest point detectors," *International Journal of Computer Vision,* vol. 37, pp. 151–172, Feb. 2000.

[26] S. M. Harris C., "A combined corner and edge detector," in *Alvey Vision Conf.* vol. 1, 1988, pp. 147-151.

[27] A. F. Gui V., Caleanu C., Fuiorea D, "Motion segmentation and analysis in video segmentation," *WSEAS Transaction on Circuits and System,* vol. 6, pp. 142-148, 2007.

[28] G. V. Ianasi C., Alexa F., Toma C., "Noncauzal Adaptative model-traking estimation for background substraction in video surveillance.," *WSEAS Transactions on Signal Processing (WSEAS Journal),* vol. 2, pp. 52-59, 2006.

[29] W. Skarbek, Koschan, A. , "Colour image segmentation – a survey.," vol. Tech. rep., Institute for Technical Informatics, Technical University of Berlin, , 1994.

[30] M. P. Judd D., Jain A.K. , "Large-Scale Parallel Data Clustering,," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, pp. 871-876, 1998.

[31] D. J. S. Bhatia S.K., "Conceptual Clustering in Information Retrieval," *IEEE Trans. Systems, Man, and Cybernetics,* vol. 28, pp. 427-436, 1998.

[32] K. R. Frigui H., "A Robust Competitive Clustering Algorithm with Applications in Computer Vision," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, pp. 450-465, 1999.

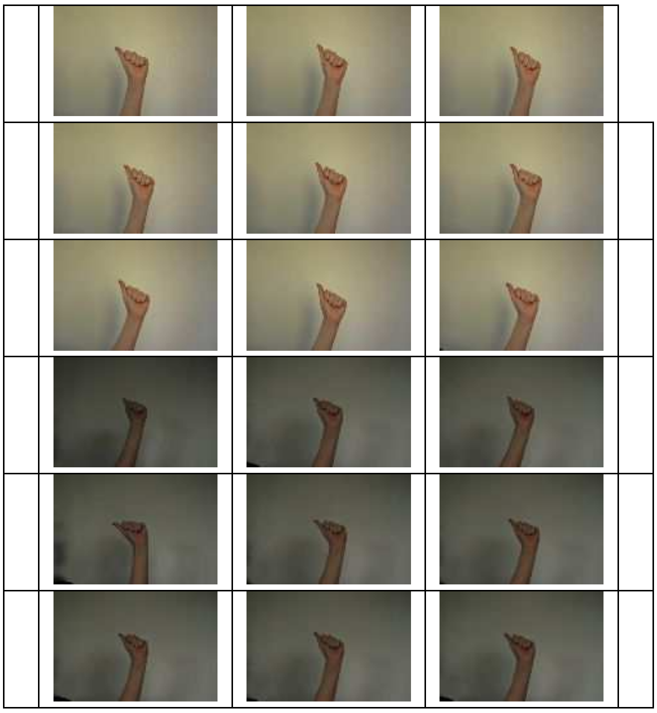
[33] F. M. M. Abbas H.M., "Neural Networks for Maximum Likelihood Clustering," *Signal Processing,* vol. 36, pp. 111-126, 1994.

[34] R. G. Carpineto C., "A Lattice Conceptual Clustering System and Its Application to Browsing Retrieval," *Machine Learning,* vol. 24, pp. 95-122, 1996.

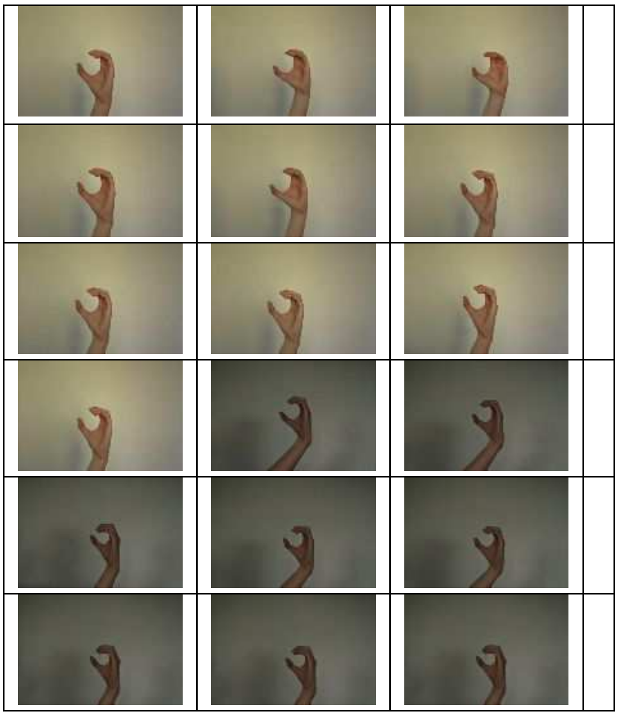
[35] Anil K. Jain, Robert P.W. Duin, Jianchang Mao, "Statistical Pattern Recognition: A Review," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE,* vol. 22, 2000.

[36] Ommer Björn, "Learning the Compositional Nature of Objects for Visual Recognition," *Phd Thesis* 2007.

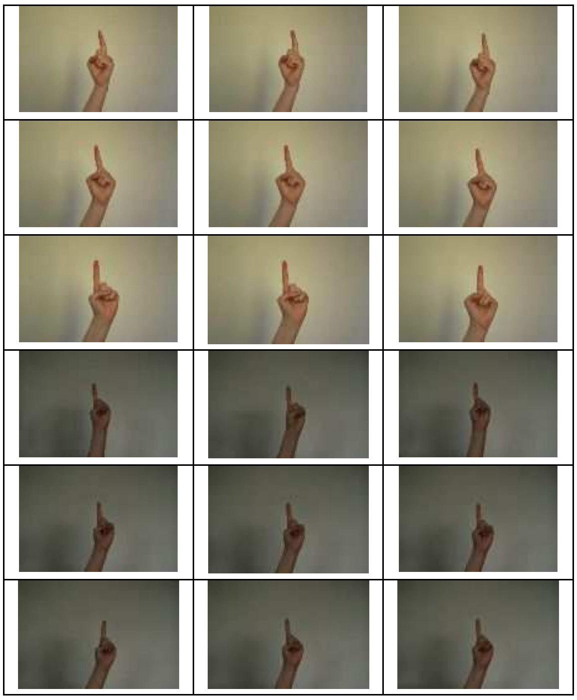
Appendix



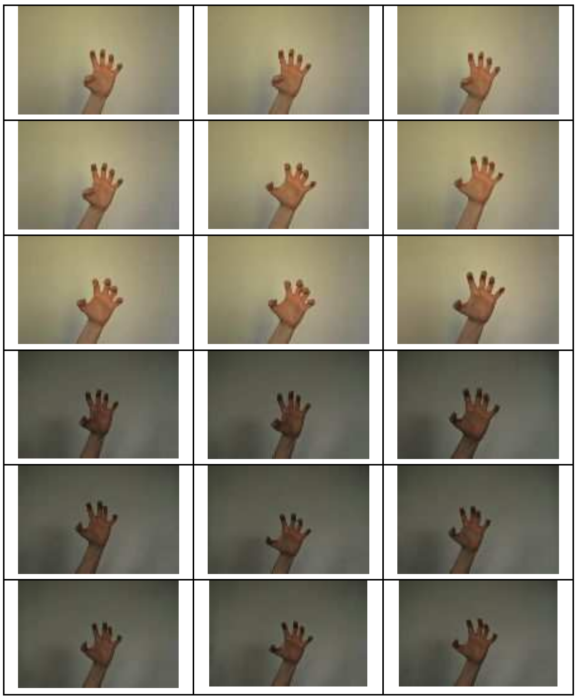
Gesture a



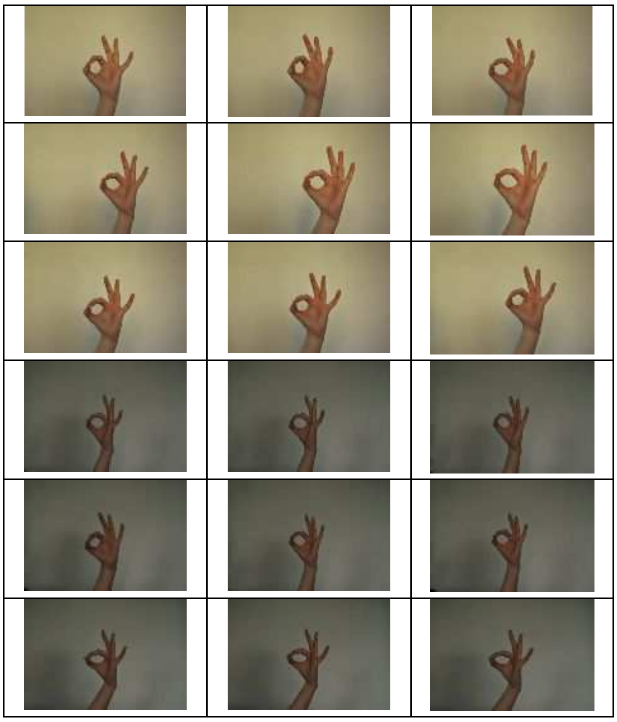
Gesture c



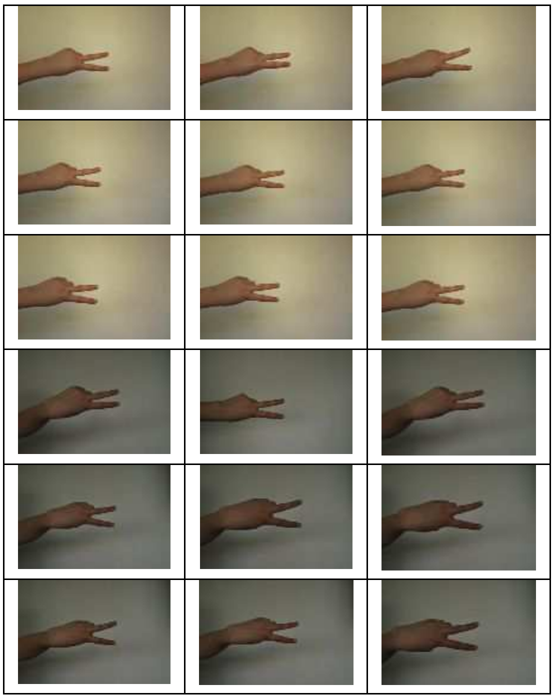
Gesture d



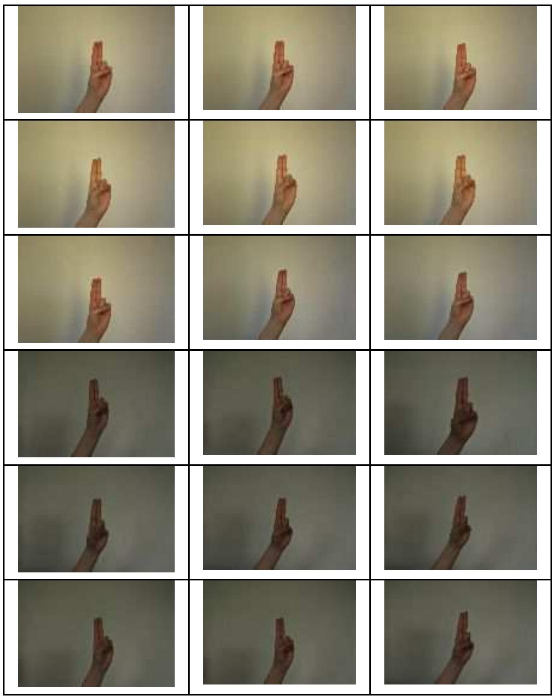
Gesture e



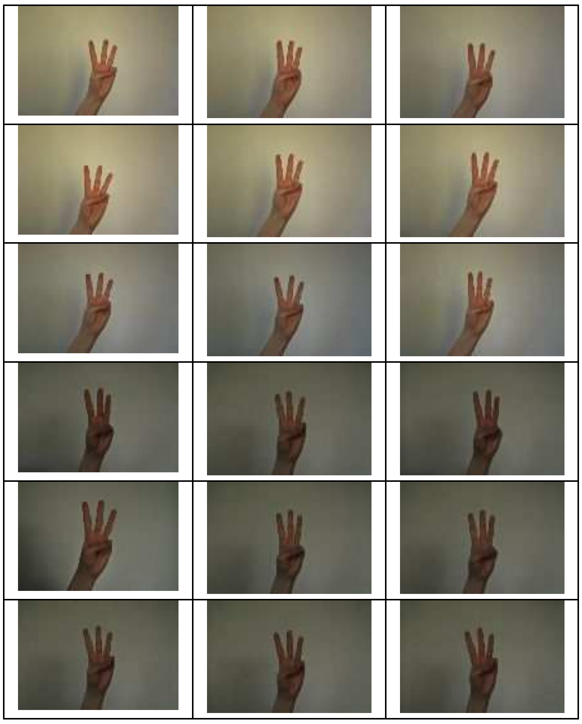
Gesture f



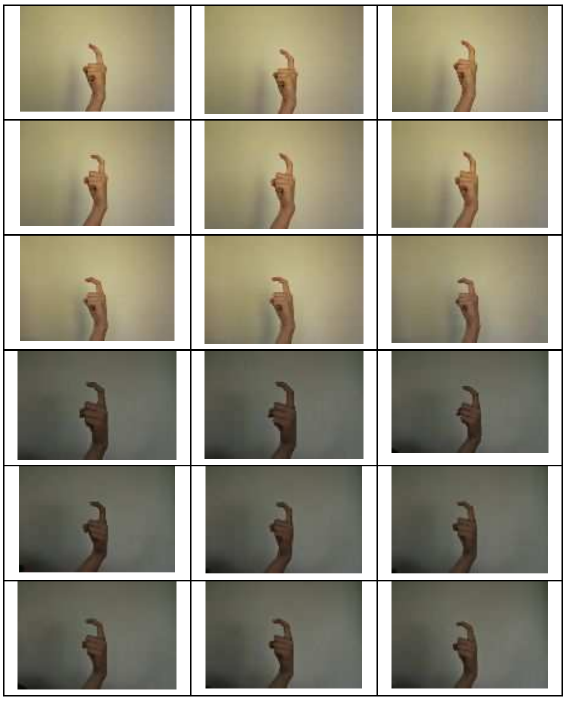
Gesture p



Gesture u



Gesture w



Gesture x