A New Compositional Technique for Hand Posture Recognition
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: - A new compositional technique for hand posture recognition is described. 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.

Learning object representations for detection and recognition is a challenging task in computer vision. The recent vision literature has observed a trend for returning to the compositional and grammatical methods, for example, the work in the groups of Ahuja [1], Geman [2, 3], Dickinson [4, 5], Pollak [6], Buhmann [7] and Zhu [8, 9, 10, 11].

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 hand model [12, 13, 14, 15].

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 obviously problems show up when skin like objects are in scene. Zhou [16] 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 [17, 18, 19, 20].

We propose an alternative approach for hand pose recognition based on compositional techniques. The compositional techniques have very good results for object categorization when we deal with many classes and a high degree of variability inside the class. Fei- Fei and Perona [21] used this technique to recognize natural scene categories. Fergus [22] learned object categories from Google’s image search. Sparse representations are compositional techniques. Patches, atoms, salient points, interest points, edges [7, 23, 24] are representatively features for sparse representation.

In this paper a method for learning the compositional structure of hand posture is proposed, and the result is presented in a gesture recognition system. Characteristic compositions of atomic parts are learned for each hand posture in an unsupervised manner, requiring no hand segmentations or localization during training. The salient points from the image are detected using Harris interest point detector. Color and edge histogram over a small patches are computed. A small codebook from a set of atomic parts that are shared among categories is generated.

The paper is organized as follows: second chapter introduces related works and the outline of our method; the third chapter describes our method in details and experiments; the fourth chapter includes the conclusion of our method.
2 Theoretical Background

Human vision is a complex process. According to Attnave [25] the visual stimulus is highly redundant in the sense that there are significant spatial dependencies in visual scenes. It is proved that people can recognize complex entities objects by means of comparably few, simple, and widely usable parts together with relations between them. The compositional approach are based on this fact and learns intermediate groupings of parts and builds a bridge over the semantic gap between low level features and high level object recognition by establishing intermediate hidden layer representations. The compositional techniques have very good results for object categorization when we deal with complex applications from natural scene categorization to Google’s image search categorization [21, 22].

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 [26], SIFT features [27], geometric blur, Gabor filters, localized histograms [7,28] 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 [26] are single layered. Recently, probabilistic latent semantic analysis (pLSA) [29] and latent Dirichlet allocation [30] are used to introduce a hidden representation layer of abstract concepts [31], [22].

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.

Degree of Supervision: The restriction of user assistance is desirable or scaling methods up to large numbers of categories with large training sets.

3 Problem Solution

In our approach localized features histograms are used. The RGB image is converted to a gray scale image, and then Canny edge detector is used in order to extract the hand contours. The salient image locations are detected using Harris interest point detector on hand contours, and quadratic patches of size $20 \times 20$ pixels are extracted. For each extracted patch its correspondent in the RGB image is searched and a two bin color histogram (skin-non skin) is extracted. The contour direction histogram with four bins and the relative direction of the interest point are also extracted from the same image patch. Patches, atoms, salient points, interest points, edges [15, 23, 24] are representatively features for sparse representation. All seven parameters extracted from a patch are used to form a feature vector $e_1$. It is important to remark the small dimension of the feature vector, which is seven. 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.

In order to have a codebook with relevant features mean shift clustering over the feature vectors is used. The mean shift algorithm is a nonparametric clustering technique which does not require 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 mean shift vector and translation of the window is guaranteed to converge to a point where the gradient of density function is zero.

$$m_h(x) = \frac{\sum_{i=1}^{n} g \left( \frac{|x-x_i|}{h} \right)^2 }{\sum_{i=1}^{n} g \left( \frac{|x-x_i|}{h} \right)^2} - x$$

(1)

The mean shift clustering algorithm is a practical application of the mode finding procedure: starting on the data points than mean shift procedure to find the stationary points of the density function is run. Currently the number of prototypes for this application is 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 have: 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, or the patch may have only background respectively skin regions. To robustify the representation, each feature point is described by a Gibbs distribution over the codebook like in [28] instead of its nearest prototype:

\[
P(F_i = ν | e_i) = Z(\mathbf{e}_i)^{-1} \exp(-d_{ν,σ}(\mathbf{e}_i))
\]

\[
Z(\mathbf{e}_i) = \sum_{ν} \exp(-d_{ν,σ}(\mathbf{e}_i))
\]

Where \(F_i\) is a feature assignment random variable, \(d_{ν,σ}(\mathbf{e}_i)\) is the Euclidian distance of a measured feature \(e_i\) to a centroid \(a_ν\) of class \(ν\) and \(σ\) is a normalization factor. Eq. (2) is evaluated for all centroids \(a_ν\) and the results for a feature point are grouped in a part distribution vector \(d_i = (P(F_1 = 1 | e_i), ..., P(F_k = k | e_i))^T\).

The framework diagram for the training step can be seen in Fig. 1.

In the following step, based on the principle of perceptual organization (the grouping principle of proximity from Gestalt laws), all detected local parts from an image are grouped with their neighbors which are not farther away than 25 pixels. Candidate compositions are represented as mixtures of the part (feature point) distributions as defined in Eq. (2). If \(\Gamma_j = \{e_1,...,e_m\}\) denotes the grouping of parts represented by \(e_1,...,e_m\), 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:

\[
g_j = \frac{1}{m} \sum_{i=1}^{m} d_i = \frac{1}{m} \sum_{i=1}^{m} (P(F_1 = 1 | e_i), ..., P(F_k = k | e_i))^T
\]

On the set of all compositions formed, a selection must be performed. The relevant compositions must reflect a trade-off between generality and singularity. First, compositions which are specific for a large majority of classes are learned. These compositions should be shared by many categories. In order to do this, in the learning phase, mean shift clustering is used over all compositions formed from the set of training images. Then the prior assignment probabilities of candidates to clusters \(P(\pi_1)\), \(\pi_1 \in \Π\), are computed using Gibbs distribution according to Eq. (2) and Eq. (3). In the second stage, relevant prototypes for a class are selected. Those prototypes help to distinguish between classes. The category posteriors of compositions must be estimated.

![Fig. 1 The framework diagram for the training step](image)

In order to estimate the category posteriors of compositions we used a Bayesian approach:

\[
P(c | \Gamma_j) = \frac{P(\Gamma_j | c)P(c)}{P(\Gamma_j)} = \frac{P(\Gamma_j | c)P(c)}{\sum_c P(\Gamma_j | c)P(c)}
\]

\[
P(c | \Gamma_j) = \frac{P(\Gamma_j | c)}{\sum_c P(\Gamma_j | c)}
\]

where \(c \in φ\), \(φ\) is the set of all category labels. The category posterior is used to calculate the relevance of a composition for discriminating categories. The entropy is used as a measure of discriminative relevance.
\[ H(P_j) = - \sum_{c \in C_j} P(c|\Gamma_j) \log P(c|\Gamma_j) \]  

(6)

In order to measure the total relevance of a prototype, a cost function is defined. This function combines the prior assignment probabilities of clusters and entropy.

\[ S(\pi_i) \propto -P(\pi_i) + \lambda H(P_{\pi_i}) \]  

(7)

A set of relevant prototypes was selected. The recognition part is done with the bag of features method and the framework diagram is shown in Fig. 2.

![Fig. 2 The framework diagram for recognition part with the bag of features method](image)

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 compositions. In order to get invariance to translation the relative coordinates are used. The hand position is estimated using the median, not the mean because the median is more robust.

\[
\mathbf{h}_i = \begin{bmatrix} x_i \\ y_i \\ d_i \end{bmatrix}
\]

(8)

where \( x_i = \alpha(x - x_{\text{median}}) \), \( y_i = \alpha(y - y_{\text{median}}) \)

![Fig. 3 Hand postures](image)

The set of vectors is compared with the set of vectors computed for all training images. The minimum distance between the set of vectors for the new image and all trained images discriminates the hand posture.

In this paper nine targeted hand posture classes, named “first”, “second”, “third”, “Ok”, “fourth”, “two fingers”, “fifth”, “one finger”, “three fingers” are trained and recognized. The hand posture had limited variation. The hand postures set can be seen in Fig. 3. The total number of images in test set was 135, for each hand posture we had 15 samples. The
pictures were taken in normal illumination conditions (daylight, no artificial light was added) and white wall as background. Our images had the resolution 256 x 171. Given the small number of samples per class we tested our performances using the live one out method.

In our experiments the current number of prototypes, was set to 20 and the number of relevant prototypes \( r \) to 19. At this point the parameter \( \sigma \) from Eq.(2) and Eq.(3) was set to 0.05, this parameter influences the selectivity. The parameter \( \lambda = 3 \) was used with best results and reported in the paper. For all sets of composition vectors computed with Eq. (8) the relative position was scaled with factor \( \alpha = 0.1 \).

The confusion matrix can be seen in Fig. 4.

We consider our results to be promising; we are aware of the small number of samples, our future plan is to increase our data base. Until now, only a limited number of optimizations have been done. A challenging task is to find features that help discriminating among an extended set of hand postures.

Our results compare favorably with other results reported in literature. The error rate reported in [8] was 0.02 for posture “palm”, 0.029 for posture “little finger”, 0.02 for “first” and 0.025 for “two–finger”. In these experiments different illumination conditions were used and white wall was used as background. In [26] there are 8 static gestures, the error rate was 0.25 for frontal view of hand posture and 0.396 for back view of hand posture (the result was ranked within first). In [5] there are three hand postures “palm”, “first”, and “sixth”, the error rate was 0.121, 0.041 and 0.117 respectively.

4 Conclusion

In this paper a compositional approach for hand posture recognition is proposed. A hand posture is decomposed into relevant compositions which are learned for each class of gestures without supervision; no hand segmentations or localization during training is needed. Gesture recognition is done using the bag of feature method.

The compositional techniques are usually used for large data base categorization not for classification. The main advantage of these techniques is the generality; these techniques are more independent of application.

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:


