Finger Spelling Recognition from RGB-D Information using Kernel Descriptor

K. Otiniano-Rodríguez, G. Cámara-Chávez
Department of Computer Science (DECOM)
Federal University of Ouro Preto
Ouro Preto, MG, Brazil
Email: {karlaotiniano,gcamarac}@gmail.com

***Abstract*—Deaf people use systems of communication based on sign language and finger spelling. Manual spelling, or finger spelling, is a system where each letter of the alphabet is represented by an unique and discrete movement of the hand. RGB and depth images can be used to characterize hand shapes corresponding to letters of the alphabet. The advantage of depth cameras over color cameras for gesture recognition is more evident when performing hand segmentation. In this paper, we propose a hybrid system approach for finger spelling recognition using RGB-D information from Kinect™ sensor. In a first stage, the hand area is segmented from background using depth map and precise hand shape is extracted using both depth data and color data from Kinect™ sensor. Motivated by the performance of kernel based features, due to its simplicity and the ability to turn any type of pixel attribute into patch-level features, we decided to use the gradient kernel descriptor for feature extraction from depth images. The Scale-Invariant Feature Transform (SIFT) is used for describing the content of the RGB image. Then, the Bag-of-Visual-Words approach is used to extract semantic information. Finally, these features are used as input of our Support Vector Machine (SVM) classifier. The performance of this approach is quantitatively and qualitatively evaluated on a dataset of real images of American Sign Language (ASL) hand shapes. Three experiments were performed, using a combination of RGB and depth information and also using only RGB or depth information separately. The database used is composed of 120,000 images. According to our experiments, our approach has an accuracy rate of 91.26% when RGB and depth information is used, outperforming other state-of-the-art methods.**

***Keywords*-sign language; finger spelling; support vector machine (SVM); bag-of-visual-words.**

I. Introduction

Sign language is a complex way of communication in which hands, limbs, head, facial expression and body language are used to communicate a visual-spatial language without sound, mostly used between deaf-mute people. Deaf people use systems of communication based on sign language and finger spelling. In sign language, the basic units are composed by a finite set of hand configurations, spatial locations, and movements. Their complex spatial grammars are remarkably different from the grammars of spoken languages [1], [2]. Hundreds of sign languages, such as ASL (American Sign Language), BSL (British Sign Language), Auslan (Australian Sign Language) and LIBRAS (Brazilian Sign Language) [1], are in use around the world and are at the cores of local deaf cultures. Unfortunately, these languages are barely known outside of the deaf community, meaning a communication barrier.

Manual spelling, or finger spelling, is a system where each letter of the alphabet is represented by a unique and discrete movement of the hand. The finger spelling integrates a sign language due to many reasons: when a concept lacks a specific sign, for proper nouns, for loan signs (signs borrowed from other languages), for finger spelled compounds or when a sign is ambiguous [3]. Each sign language has its own finger spelling similar to different characters in different languages.

Several techniques have been developed to achieve an adequate recognition rate of sign language. Over the years and with the advance of technology, methods have been proposed in order to improve the data acquisition, processing or classification, such is the case in image acquisition. There are three main approaches: sensor-based, vision-based and hybrid systems using a combination of these systems. Sensor-based methods use sensory gloves and motion tracker to detect hand shapes and body movements. Vision-based methods, that use standard cameras, image processing, and feature extraction, are used for capturing and classifying hand shapes and body movements. Hybrid systems use information from vision-based camera and other type of sensors like infrared depth sensors.

Sensor-based methods, such as data gloves, can provide accurate measurements of hands and movement. Unfortunately, these methods require extensive calibration, they also restrict the natural movement of hands and are often very expensive. Video-based methods are less intrusive, but new problems arise: locating the hands and segmenting them is a non-trivial task. Recently, depth cameras have become popular at a commodity price. Depth information makes the task of segmenting the hand from the background much easier. Depth information can be used to improve the segmentation process, as used in [4], [5], [6], [7].

Recently, depth cameras have raised a great interest in vision computer community due to their success in many applications, such as pose estimation [8], [9], tracking [10], object recognition [10], etc. Depth cameras were also used for hand gesture recognition [11], [12], [13]. Uebersax et al. [12] present a system for recognizing letter and finger spelled words. Pugeault & Bowden [11] use a Microsoft

Kinect™ device to collect RGB and depth images. They extracted features using Gabor filters and then a Random Forest predicts the letters from the American Sign Language (ASL) finger spelling alphabet. Issacs & Foo [14] proposed an ASL finger spelling recognition system based on neural networks applied to wavelets features. Bergh & Van Gool [15] propose a method based on a concatenation of depth and color-segmented images, using a combination of Haar wavelets and neural networks for 6 hand poses recognition of a single user.

In this paper, we propose a framework for finger spelling recognition using RGB and depth images. Motivated by the performance of kernel based features, due to its simplicity and the ability to turn any type of pixel attribute into patch-level features, we decided to use the gradient kernel descriptor[16]. The experiments are performed using a public database composed of 120,000 images stating 24 symbols classes [17]. The obtained results show that the accuracy obtained by our method, using RGB and depth images, is greater than only using RGB or depth images separately. Moreover, the accuracy obtained by the proposed method performs better than the method proposed in [11]. The results show that our method is promising.

The remainder of this paper is organized as follows. In Section II, our proposed method is introduced and detailed. The experiments are presented in Section III, where the results are discussed. Finally, conclusion and future work are presented in Section IV.

II. Proposed Model

This section describes the methodology developed to perform finger spelling recognition from RGB and depth information. The proposed model consists of four stages as shown in Figure 1. In the first stage, the hand area is segmented from background using depth map and precise hand shape is extracted using both depth data and color data from Kinect™ sensor. The second stage consists in extracting the features from intensity and depth images. The SIFT descriptor is used to extract features from the intensity image. First, interest points are detected and described by its neighborhood. The Gradient kernel descriptor is used on the depth image. It consists of three kernels. The normalized linear kernel weighs the contribution of each pixel using gradient magnitudes, an orientation kernel computes the similarity of gradient orientations and finally a position Gaussian kernel measures how close two pixels are spatially. The third stage consists in capturing the semantic information. In order to do this, the Bag-of-Visual-Words model is applied. Finally, these features are used as input to our SVM classifier.

A. Segmentation

The segmentation can easily be performed on depth values using a threshold, depth values corresponding to the hand are the smallest, this means that they are closer to the sensor. The segmented hand (binary image) is used as a mask over the depth and intensity image to only get the hand values.

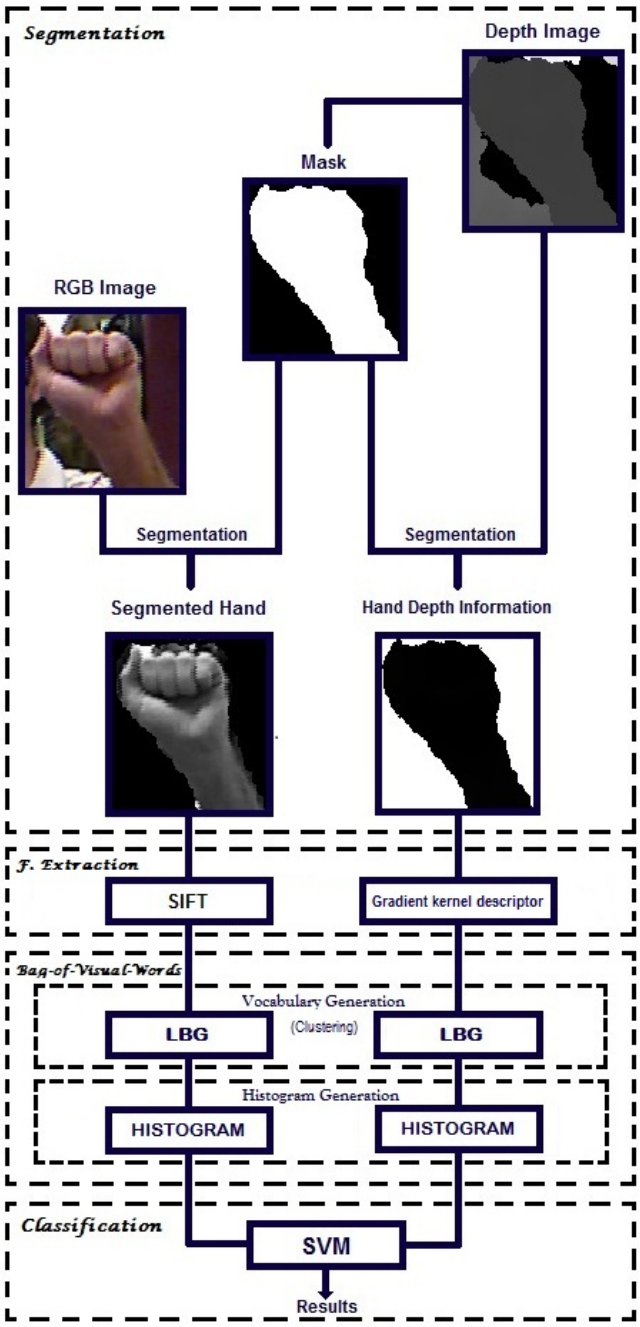

Fig. 1. Proposed model for finger spelling recognition.

Because the depth and RGB images are not aligned, an additional process is performed. The opening morphological operation is applied over the mask before segmenting the intensity image using a square structuring element with 3 pixels of width.

B. Feature extraction

With the segmented images already obtained, the next step consists of extracting the features. In order to do this, gradient

kernel descriptor and SIFT are applied over the depth image and intensity image, respectively.

1) Gradient kernel descriptor: The low-level image feature extractor, kernel descriptor, designed for visual recognition in [18], consists of three steps: design match kernel using some pixel attribute, learn compact basis vectors using Kernel Principle Component Analysis and construct kernel (KPCA) descriptor by projecting the infinite-dimensional feature vector to the learned basis vectors. The authors leading three types of effective kernel descriptors using gradient, color and shape pixel attributes. In other model proposed by the same authors [16], the gradient kernel descriptor is applied over depth images. Thereby, in order to capture edge cues in depth maps, we used the gradient match kernel, K_{grad} :

$$K_{grad}(P,Q) = \sum_{p\in P}\sum_{q\in Q} \tilde{m}(p)\tilde{m}(q)k_o(\tilde{\theta}(p),\tilde{\theta}(q))k_s(p,q)$$

The normalized linear kernel $\tilde{m}(p)\tilde{m}(q)$ weighs the contribution of each gradient where $\tilde{m}(p) = m(p)/\sqrt{\sum_{p\in P} m(p)^2 + \varepsilon_g}$ and ε_g is a small positive constant to ensure that the denominator is larger than 0 and $m(p)$ is the magnitude of the depth gradient at a pixel p. Then, $k_o(\tilde{\theta}(p),\tilde{\theta}(q)) = exp(-\gamma_o\|\tilde{\theta}(p) - \tilde{\theta}(q)\|^2)$ is a Gaussian kernel over orientations. The authors [18] suggest to set $\gamma_o = 5$. To estimate the difference between orientations at pixels p and q, we use the following normalized gradient vectors in the kernel function k_o:

$$\tilde{\theta}(p) = [sin(\theta(p))cos(\theta(p))]$$
$$\tilde{\theta}(q) = [sin(\theta(q))cos(\theta(q))]$$

where $\theta(p)$ is the orientation of the depth gradient at a pixel p. Gaussian position kernel $k_s(p,q) = exp(-\gamma_s\|p - q\|^2)$ with p denoting the 2D position of a pixel in an image patch (normalized to [0,1]), measures how close two pixels are spatially. The value suggest for γ_s is 3.

To summarize, the gradient match kernel K_{grad} consists of three kernels: the normalized linear kernel weighs the contribution of each pixel using gradient magnitudes; the orientation kernel k_o computes the similarity of gradient orientations; and the position Gaussian kernel k_s measures how close two pixels are spatially.

Match kernels provide a principled way to measure the similarity of image patches, but evaluating kernels can be computationally expensive when image patch are large [18]. The corresponding kernel descriptor can be extracted from this match kernel by projecting the infinite-dimensional feature vector to a set of finite basis vectors, which are the edge features that we use in the next steps. For more details, the approach to extract the compact low-dimensional features from match kernels is found in [18].

2) Scale-invariant feature transform (SIFT): Is an algorithm useful in computer vision to detect and describe local features in images[19].

The SIFT descriptor firstly detects interest points by scale-space extreme of Differences-of-Gaussians (DoG) within a DoG pyramid. Then the position-dependent histograms of local gradient directions around the interest points are statistically accumulated as the SIFT descriptor. In the end, this SIFT descriptor is utilized to match corresponding interest points between different images [20].

In one image can exist different objects, for any of those objects, interesting points can be extracted to provide a feature descriptor of the object. This descriptor can then be used to identify the object when attempting to locate the object in another image containing many other objects.

An important characteristic of these features is that the relative positions between them in the original scene should not change from one image to another. For example, if only the four corners of a door were used as features, they would work regardless of the door's position; but if points in the frame were also used, the recognition would fail if the door is opened or closed. Similarly, features located in articulated or flexible objects would typically not work if any change in their internal geometry happens between two images in the set being processed. However, in practice SIFT detects and uses a much larger number of features from the images, which reduces the contribution of the errors caused by these local variations in the average error of all feature matching errors.

The SIFT is invariant to translation, rotation and scaling transformations. The large number of features in a typical image allow for robust recognition under partial occlusion in cluttered images.

C. Bag-of-Visual-Words

Bag-of-Visual-Words has first been introduced by Sivic for video retrieval [21]. Due to its efficiency and effectiveness, it became very popular in the fields of image retrieval and categorization. Image categorization techniques rely either on unsupervised or supervised learning.

Our model uses the Bag-of-Visual-Words approach in order to capture semantic information. The original method works with documents and words. Therefore, we consider an image as a document and the "words" will be the visual entities found in the image. The Bag-of-Visual-Words approach consists of three operations: feature description, visual word vocabulary generation and histogram generation. In our case, the local descriptor SIFT and Gradient kernel descriptor are used in the feature description step. Then, a visual word vocabulary is generated from the feature vectors, each visual word (codeword) represents a group of several similar features. The visual word vocabulary (codebook) defines a space of all entities occurring in the image. Finally, a histogram of visual words is created by counting the occurrence of each codeword. These occurrences are counted and arranged in a vector. Each vector represents the features for an image.

D. Classification

Support vector machines, introduced as a machine learning method by Cortes and Vapnik [22], are a useful classification

method. Furthermore, SVMs have been successfully applied in many real world problems and in several areas: text categorization, handwritten digit recognition, object recognition, etc. The SVMs have been developed as a robust tool for classification and regression in noisy and complex domains. SVM can be used to extract valuable information from data sets and construct fast classification algorithms for massive data.

An important characteristic of the SVM classifier is to allow a non-linear classification without requiring explicitly a non-linear algorithm thanks to kernel theory.

In kernel framework data points may be mapped into a higher dimensional feature space, where a separating hyperplane can be found. We can avoid to explicitly compute the mapping using the kernel trick which evaluate similarities between data $K(d_t, d_s)$ in the input space. Common kernel functions are: linear, polynomial, Radial Basis Function (RBF), χ^2 distance and triangular.

III. Experiments

The ASL Finger Spelling Dataset [17] contains 500 samples for each of 24 signs, recorded from 5 different persons (non-native to sign language), amounting to a total of 60,000 samples. Each sample has a RGB image and a depth image, making a total of 120,000 images. The sign J and Z are not used, because these signs have motion and the proposed model only works with static signs. The dataset has variety of background and viewing angles. Figure 2 shows some examples. It is possible to see the variety in size, background and orientation.

Due to the variety in the orientation when the signal is performed, signs became strongly similar. Figure 3 shows the most similar signs *a*, *e*, *m*, *n*, *s* and *t*. The examples are taken from the same user. It is easy to identify the similarity between these signs, all are represented by a closed fist, and differ only by the thumb position, leading to higher confusion levels. Therefore, these signs are the most difficult to differentiate in the classification task.

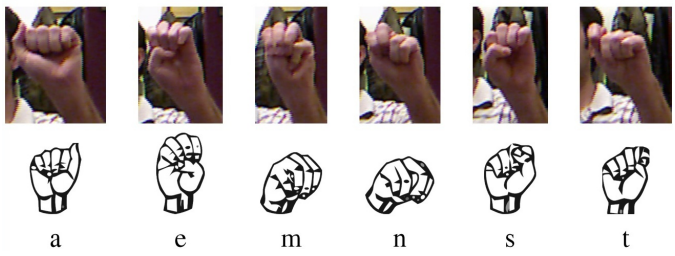

Fig. 3. Most conflictive similar signs in the dataset.

In order to validate our technique, we conduct three experiments. In the first, a classification of the signs was performed using only the RGB image features. In the second, a classification was performed using the depth image features. Finally, in the third experiment the signs were classified using both features (RGB-D). For each experiment, we have some specifications. To extract all low level features using gradient kernel descriptor, are used approximatively 12x13 patches over dense regular grid with spacing of 8 pixel (images are not of uniform size). In order to produce the visual word vocabulary, the LBG (Linde-Buzo-Gray) [23] algorithm was used to detect one hundred clusters by taking a sample of 30% from the total features. Moreover, in the classification stage, we use a RBF kernel, whose values for g (gamma) and c (cost) are 0.25 and 5, respectively. We also use a cross validation with 5 folds. The library LIBSVM (a library for Support Vector Machines)] [24] was used in our implementation.

First experiment: An average accuracy of 63% was obtained. This accuracy is the mean of the values of the main diagonal of the confusion matrix and represents the signs correctly classified (true positives). The confusion matrix for this experimente is found in table I, here, we can observe that signs *n*, *r*, *k* and *x* have the lowest averages (between 50% and 54%), and the sign *h* has the highest average (82%). This shows the wide variation between the results for this experiment. It means that there is not enough information. Therefore, we will use depth information in order to increase this results.

Second experiment: For this experiment, the average accuracy obtained was 86%. This result shows an increase of 23% in the recognition rate compared to the previous experiment. Signs *n*, *r*, *k* and *x* improved in 24%, 25%, 30%, and 24% respectively. Table II shows the confusion matrix for the classification task using depth information.

Third experiment: The classification task was performed using RGB-D information, obtaining an average accuracy of 91.26%. The data for this experiment was obtained by joining the features (histograms) from RGB and depth information, which were used in the experiments 1 and 2, respectively. Table III shows the results for this experiment. Signs *b*, *c*, *f*, *i*, *l* and *y* have the highest average accuracies (over 95%). Otherwise, the signs *n*, *r* and *t* have the lowest values (between 82% and 84%). The low recognition value of sign *n* is due to the big similarity with signs *m* and *t*, as shown in the Figure 3. An improvement was obtained over these signs compared to the results of experiments 1 and 2. The sign *n* had an accuracy of 82%, it means 32% and 8% improvement over the first and second experiment, respectively.

We summarize and compare the results in table IV. It includes the average accuracy and standard deviation. We can see that RGB-D information obtains the highest average accuracy, outperforming the RGB and depth methods and also the method proposed by Pugeault & Bowden [11]. This last method is found in the state-of-the-art and uses the same dataset, obtaining an average accuracy of 75%. About standard deviation, we can see the low variation when RGB-D information is used.

IV. Conclusion and Future work

In this paper, we propose a method for Finger Spelling Recognition using RGB-D information, combining intensity and depth descriptors. Then, Bag-of-Visual-Words was applied

Fig. 2. ASL Finger Spelling Dataset: 24 static signs by 5 users. It is an example of the variety of the dataset. This array shows one image from each user and from each letter.

TABLE I
CONFUSION MATRIX OF THE CLASSIFICATION OF 24 SIGN USING RGB INFORMATION.

	a	b	c	d	e	f	g	h	i	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y
a	0.67	0.01	0.04	0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.01	0.03	0.02	0.02	0.00	0.01	0.01	0.02	0.03	0.01	0.00	0.00	0.01	0.01
b	0.02	0.74	0.02	0.01	0.01	0.02	0.00	0.00	0.02	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.02	0.01	0.01	0.03	0.02	0.01	0.01	0.01
c	0.03	0.02	0.67	0.01	0.01	0.01	0.02	0.01	0.01	0.02	0.02	0.01	0.01	0.03	0.02	0.01	0.02	0.02	0.00	0.02	0.02	0.01	0.02	0.02
d	0.01	0.02	0.02	0.55	0.03	0.03	0.01	0.01	0.02	0.02	0.02	0.01	0.02	0.02	0.02	0.01	0.04	0.02	0.02	0.02	0.02	0.02	0.02	0.03
e	0.02	0.01	0.02	0.04	0.58	0.02	0.01	0.00	0.02	0.01	0.02	0.03	0.03	0.02	0.03	0.02	0.01	0.02	0.03	0.01	0.01	0.01	0.03	0.02
f	0.00	0.01	0.01	0.02	0.02	0.72	0.00	0.01	0.03	0.01	0.01	0.00	0.01	0.00	0.01	0.00	0.02	0.00	0.00	0.01	0.01	0.05	0.01	0.02
g	0.01	0.00	0.02	0.01	0.01	0.02	0.71	0.07	0.01	0.03	0.01	0.00	0.01	0.02	0.02	0.02	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01
h	0.00	0.00	0.01	0.01	0.01	0.00	0.07	0.82	0.00	0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01
i	0.02	0.02	0.02	0.02	0.02	0.03	0.01	0.00	0.66	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02
k	0.02	0.01	0.03	0.03	0.02	0.02	0.03	0.01	0.04	0.54	0.03	0.01	0.01	0.01	0.01	0.01	0.03	0.01	0.01	0.02	0.03	0.02	0.02	0.02
l	0.02	0.01	0.02	0.02	0.01	0.01	0.00	0.00	0.02	0.02	0.73	0.00	0.01	0.01	0.01	0.00	0.02	0.01	0.01	0.02	0.01	0.00	0.01	0.04
m	0.04	0.01	0.01	0.01	0.03	0.01	0.01	0.00	0.02	0.00	0.00	0.59	0.06	0.04	0.02	0.03	0.00	0.04	0.05	0.00	0.00	0.00	0.01	0.00
n	0.04	0.02	0.01	0.02	0.04	0.01	0.01	0.00	0.03	0.01	0.01	0.07	0.50	0.02	0.02	0.02	0.01	0.03	0.06	0.01	0.00	0.01	0.03	0.01
o	0.02	0.01	0.04	0.02	0.02	0.01	0.01	0.01	0.03	0.01	0.01	0.04	0.03	0.57	0.02	0.03	0.01	0.04	0.02	0.01	0.01	0.00	0.01	0.01
p	0.01	0.01	0.01	0.02	0.03	0.01	0.01	0.00	0.01	0.01	0.01	0.02	0.02	0.02	0.67	0.07	0.01	0.00	0.02	0.01	0.00	0.00	0.02	0.01
q	0.01	0.02	0.01	0.01	0.03	0.01	0.02	0.01	0.01	0.01	0.01	0.02	0.02	0.03	0.08	0.63	0.01	0.01	0.02	0.01	0.00	0.01	0.01	0.01
r	0.02	0.03	0.03	0.04	0.02	0.03	0.01	0.01	0.02	0.04	0.03	0.00	0.01	0.01	0.01	0.00	0.51	0.01	0.00	0.06	0.06	0.03	0.02	0.02
s	0.04	0.01	0.04	0.02	0.03	0.00	0.01	0.00	0.01	0.01	0.02	0.03	0.03	0.04	0.01	0.01	0.01	0.59	0.02	0.01	0.00	0.00	0.02	0.01
t	0.04	0.02	0.01	0.02	0.03	0.01	0.01	0.00	0.02	0.01	0.01	0.04	0.06	0.02	0.03	0.03	0.01	0.03	0.56	0.01	0.01	0.00	0.01	0.01
u	0.01	0.04	0.04	0.02	0.01	0.02	0.00	0.00	0.02	0.02	0.02	0.00	0.01	0.01	0.01	0.00	0.06	0.01	0.01	0.59	0.06	0.03	0.01	0.01
v	0.01	0.02	0.03	0.02	0.01	0.03	0.01	0.00	0.01	0.03	0.02	0.00	0.00	0.01	0.00	0.00	0.07	0.00	0.00	0.06	0.58	0.07	0.01	0.01
w	0.00	0.01	0.01	0.02	0.01	0.05	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.02	0.06	0.72	0.00	0.01
x	0.02	0.01	0.03	0.03	0.03	0.01	0.01	0.00	0.02	0.03	0.02	0.01	0.03	0.01	0.04	0.01	0.03	0.02	0.02	0.02	0.01	0.01	0.54	0.03
y	0.01	0.01	0.03	0.03	0.02	0.03	0.01	0.00	0.03	0.02	0.05	0.00	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.02	0.02	0.01	0.02	0.59

TABLE IV
ACCURACIES AND STANDARD DEVIATION OF THE THREE EXPERIMENTS.

Method	Accuracy	Standard Deviation
RGB	62.70%	0.47
Depth	85.18%	0.16
RGB-D	91.26%	0.18
Pugeault & Bowden[11]	75.00%	-

in order to capture the semantic information. Finally, the classification task is performed by a SVM.

The combination of RGB and depth descriptors obtain the best results (91.26%) with a low variance. Our method achieves a better differentiation of similar signs like *n*, *r* and *t*, incrementing the recognition rate. The Gradient kernel descriptor has the advantage that can be directly applied on the depth images without having to compute the cloud of points, consequently, reducing the computation time.

As future work, we pretend to test other kernels over depth images. We also intend to extend our method to recognize dynamic signs.

TABLE II
CONFUSION MATRIX OF THE CLASSIFICATION OF 24 SIGN USING DEPTH INFORMATION.

	a	b	c	d	e	f	g	h	i	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y
a	0.89	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.01	0.05	0.00	0.00	0.00	0.01	0.00
b	0.01	0.95	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
c	0.01	0.00	0.92	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
d	0.00	0.00	0.00	0.87	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.02	0.00	0.00	0.02	0.00	0.01	0.01	0.00	0.00	0.01	0.00
e	0.02	0.00	0.02	0.01	0.82	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.04	0.00	0.00	0.00	0.02	0.01	0.00	0.00	0.00	0.01	0.00
f	0.00	0.02	0.01	0.01	0.00	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
g	0.01	0.00	0.00	0.00	0.00	0.00	0.86	0.09	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
h	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
i	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.94	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
k	0.00	0.00	0.00	0.04	0.00	0.00	0.02	0.00	0.00	0.84	0.01	0.00	0.00	0.00	0.01	0.00	0.02	0.00	0.01	0.01	0.01	0.00	0.01	0.00
l	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
m	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.77	0.07	0.01	0.00	0.00	0.01	0.05	0.04	0.00	0.00	0.00	0.00	0.00
n	0.02	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.07	0.74	0.02	0.00	0.00	0.00	0.02	0.07	0.00	0.00	0.00	0.02	0.00
o	0.01	0.00	0.02	0.01	0.06	0.00	0.01	0.00	0.00	0.00	0.00	0.02	0.01	0.78	0.01	0.01	0.00	0.02	0.01	0.00	0.00	0.00	0.01	0.00
p	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.88	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
q	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.88	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
r	0.01	0.01	0.00	0.06	0.00	0.01	0.00	0.00	0.00	0.03	0.01	0.01	0.00	0.00	0.00	0.00	0.76	0.00	0.01	0.05	0.03	0.00	0.01	0.00
s	0.03	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.01	0.00	0.05	0.03	0.03	0.00	0.00	0.00	0.76	0.04	0.00	0.00	0.00	0.01	0.00
t	0.06	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.08	0.02	0.00	0.00	0.00	0.05	0.68	0.00	0.00	0.00	0.02	0.00
u	0.00	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.82	0.03	0.02	0.01	0.00
v	0.00	0.00	0.00	0.02	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.03	0.00	0.00	0.03	0.83	0.04	0.00	0.00
w	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.04	0.91	0.00	0.00
x	0.01	0.00	0.01	0.03	0.01	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.03	0.01	0.01	0.01	0.02	0.01	0.02	0.01	0.01	0.00	0.78	0.00
y	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.94

TABLE III
CONFUSION MATRIX OF THE CLASSIFICATION OF 24 SIGN USING RGB-D INFORMATION.

	a	b	c	d	e	f	g	h	i	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y
a	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.00
b	0.00	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
c	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
d	0.00	0.00	0.00	0.91	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00
e	0.01	0.00	0.01	0.01	0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.01	0.00
f	0.00	0.01	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
g	0.00	0.00	0.00	0.00	0.00	0.00	0.92	0.04	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
h	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
i	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.96	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
k	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.91	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.01	0.00
l	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
m	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.05	0.01	0.00	0.00	0.00	0.03	0.02	0.00	0.00	0.00	0.00	0.00
n	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.06	0.82	0.01	0.00	0.00	0.00	0.01	0.04	0.00	0.00	0.00	0.01	0.00
o	0.01	0.00	0.01	0.01	0.03	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.87	0.00	0.00	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00
p	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.93	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
q	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.93	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
r	0.01	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.84	0.00	0.00	0.03	0.02	0.00	0.00	0.00
s	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.00	0.00	0.00	0.88	0.01	0.00	0.00	0.00	0.01	0.00
t	0.03	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.05	0.01	0.00	0.00	0.00	0.02	0.82	0.00	0.00	0.00	0.01	0.00
u	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.88	0.02	0.01	0.01	0.00
v	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.03	0.90	0.02	0.00	0.00
w	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.95	0.00	0.00
x	0.01	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.02	0.01	0.01	0.00	0.01	0.00	0.01	0.01	0.00	0.00	0.87	0.00
y	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96

ACKNOWLEDGMENT

The authors are thankful to CNPq, CAPES and FAPEMIG (Projeto Universal 02292-12), Brazilian funding agencies for the support to this work.

REFERENCES

[1] LIBRAS, "Brazilian sign language," http://www.libras.org.br/, last visit: March 10, 2012.

[2] P. W. Vamplew, "Recognition of sign language gestures using neural networks," *Australian Journal of Intelligent Information Processing Systems*, vol. 5, pp. 27–33, 1996.

[3] A. Puente, J. M. Alvarado, and V. Herrera, "Fingerspelling and sign language as alternative codes for reading and writing words for Chilean deaf signers," *American Annals of the Deaf*, vol. 151, no. 3, pp. 299–310, 2006.

[4] Z. Ren, J. Yuan, and Z. Zhang, "Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera," in *Proceedings of the 19th ACM International Conference on Multimedia*. ACM, 2011, pp. 1093–1096.

[5] V. Frati and D. Prattichizzo, "Using Kinect for hand tracking and rendering in wearable haptics," in *Proceedings of the IEEE World Haptics Conference (WHC)*. IEEE, 2011, pp. 317–321.

[6] Y. Li, "Hand gesture recognition using Kinect," in *Proceedings of the 3rd IEEE International Conference on Software Engineering and Service Science (ICSESS)*. IEEE, 2012, pp. 196–199.

[7] Z. Mo and U. Neumann, "Real-time hand pose recognition using low-resolution depth images," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE, 2006, pp. 1499–1505.

[8] G. Fanelli, J. Gall, and L. V. Gool, "Real time head pose estimation with random regression forests," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011, pp. 617–624.

[9] J. Shotton, T. Sharp, A. Kipman, A. W. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[10] I. Oikonomidis, N. Kyriazis, and A. Argyros, "Efficient model-based 3D tracking of hand articulations using Kinect," in *Proceedings of the British Machine Vision Conference*. BMVA Press, 2011, pp. 101.1–101.11.

[11] N. Pugeault and R. Bowden, "Spelling it out: Real-time ASL fingerspelling recognition." in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2011, pp. 1114–1119.

[12] D. Uebersax, J. Gall, M. V. den Bergh, and L. J. V. Gool, "Real-time sign language letter and word recognition from depth data," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011, pp. 383–390.

[13] M. d. S. Anjo, E. B. Pizzolato, and S. Feuerstack, "A real-time system to recognize static gestures of brazilian sign language (libras) alphabet using kinect," in *Proceedings of the 11th Brazilian Symposium on Human Factors in Computing Systems*. Brazilian Computer Society, 2012, pp. 259–268.

[14] J. Isaacs and S. Foo, "Hand pose estimation for american sign language recognition," *36th Southeastern Symposium on System Theory*, pp. 132–136, 2004.

[15] M. Van den Bergh and L. Van Gool, "Combining RGB and ToF cameras for real-time 3D hand gesture interaction," in *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV)*, ser. WACV '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 66–72.

[16] L. Bo, X. Ren, and D. Fox, "Depth kernel descriptors for object recognition," in *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2011, pp. 821–826.

[17] R. B. Nicolas Pugeault, "ASL finger spelling dataset," http://personal.ee.surrey.ac.uk/Personal/N.Pugeault/index.php?section=FingerSpellingDataset, last visit: April 29, 2013.

[18] L. Bo, X. Ren, and D. Fox, "Kernel descriptors for visual recognition," *Advances in Neural Information Processing Systems*, vol. 7, 2010.

[19] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2. IEEE, 1999, pp. 1150–1157.

[20] S.-H. Zhong, Y. Liu, and G. Wu, "S-sift: A shorter sift without least discriminability visual orientation," *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 1, pp. 669–672, 2012.

[21] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proceedings of the Ninth IEEE International Conference on Computer Vision*. IEEE, 2003, pp. 1470–1477.

[22] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[23] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, 1980.

[24] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.