Spatial Pyramid Matching for Finger Spelling Recognition in Intensity Images

Samira Silva [†], William Robson Schwartz [‡], and Guillermo Cámara-Chávez [†]

[†] Computer Science Department, Federal University of Ouro Preto
Ouro Preto, MG, Brazil
{samirapgti,gcamarac}@gmail.com
[‡] Computer Science Department, Federal University of Minas Gerais
Belo Horizonte, MG, Brazil
{william}@dcc.ufmg.br

Abstract. Sign language is a complex way of communication mostly used for deaf people where hands, limbs, head and facial expressions are used to communicate. Finger spelling is a system where each letter of the alphabet is represented by a unique and discrete movement of the hand. In this paper, we are interested in studying the properties of the spatial pyramid matching descriptor for finger spelling recognition. This method is a simple extension of an orderless bag-of-features image representation where local features are mapped to multi-resolution histograms and compute a weighted histogram intersection. The performance of the approach is evaluated on a dataset of real images of the American Sign Language (ASL) finger spelling. We conduct experiments considering three evaluation protocols. The first uses 10% of the data as training and the remaining as test, we achieve an accuracy rate of 92.50%. The second protocol considers 50% as training data, the accuracy rate was about 97.1%. Finally, in the third protocol, we perform a 5-fold cross-validation, where we achieve an accuracy rate of 97.9%. Our method achieves the best results in all three protocols when compared to state-of-the-art approaches. In all the experiments, we also evaluate the influence of the weights of the multi-resolution histograms. They do not have a significant influence in the experimental results.

Keywords: Finger spelling recognition, Sign language recognition, Pyramidal matching

1 Introduction

Sign language is a complex way of communication mostly used for deaf people where hands, limbs, head and facial expressions are used to communicate a visual-spatial language without sound. Very few hearing people speak sign language, thus the communication between deaf and hearing people is usually done through interpreters or text writing. The communication through interpreters is often limited by the availability of an interpreter and the high cost, while the

communication by text writing is very restrictive. For example, it is very inconvenient when people is walking, standing at a distance, or when more than two people are part of the conversation.

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 of image acquisition. There are three main approaches: *i)* sensor-based, sensory gloves and motion tracker are used to detect handshapes and body movements; *ii)* vision-based, this approach uses standard cameras, image processing and feature extraction are used for capturing and classifying handshapes and body movements; and *iii)* hybrid systems, use information from vision cameras and other type of sensors like infrared depth sensors.

Vision-based approaches are less intrusive than sensor-based approaches. Issacs & Foo [1] proposed an American Sign Language (ASL) finger spelling recognition system based on neural networks applied to wavelets features. Djamila & Larabi [2] proposed a method that uses skin colour and texture attributes with neural networks. In [3], the authors use a SIFT descriptor with Hidden Markov Models. Recently, depth cameras have raised a great interest in the computer vision community due to their success in many applications. Uebersax et al. [4] present a system for recognizing letter and finger spelled words. In [5], a Microsoft Kinect$^{\text{TM}}$ device was used to collect ASL finger spelling RGB and depth images, Gabor filters and Random Forest are used to predict the letters. Bergh & Van Gool [6] propose a method based on a concatenation of depth and color-segmented images, using a combination of Haar wavelets and neural networks. A Kernel descriptor is proposed in [7] for intensity and depth images for ASL finger spelling alphabet recognition. Although, the combination of depth and intensity descriptors achieves better results, it requires more computational effort compared to only intensity descriptors.

In this paper, we are interested first in studying the properties of a well known 2D shape descriptor of intensity images used for recognizing natural scene categories, the spatial pyramid matching [8]. This method is a simple extension of an orderless bag-of-features image representation, where local features are mapped to multi-resolution histograms and compute a weighted histogram intersection. The experiments are performed using a public database composed of 120,000 images stating 24 symbols classes [9]. Even though the proposed method only uses intensity information, the results outperform other methods that combine intensity and depth information [5, 10, 7]. The robustness of the descriptor is also demonstrated by a high recognition rate, over 90%, with a small training set (10%). We also made an extensive evaluation of the influence of the dictionary size in the accuracy of the model. Even though, the weights of the histograms allow the achievement of a better recognition rate for natural scene recognition, we demonstrate that this behavior does not correspond for sign recognition. The weights of the histograms do not have any influence in the recognition rate.

2 Spatial Pyramid Matching

In this section, we are going to present the formal definition of the spatial pyramid matching method. The goal is to develop image representations that use low-level features to infer high-level semantic information about a sign image without needing to segment the image in an intermediate step.

A *spatial pyramid* is a set of features histograms computed over cells defined by a multi-level recursive image decomposition. At level zero, the decomposition is a single cell, that is similar to a standard bag of features. On the first level, the image is subdivided into four equal parts, producing four feature histograms.

This process continues until the defined maximum level is achieved. Grauman and Darrell [11] proposed a *pyramid matching* to find an approximate correspondence between two sets of vectors in a d-dimensional feature space. Two points at a fixed resolution match if they fall into the same cell of the grid. Let X and Y be two sets of vectors and let us create a sequence of grids that has resolutions $0, \ldots, L$, where each resolution l has a 2^l cells along each dimension.

The total of cells is $D{=}2^{dl}$ cells. Let H_X^l and H_Y^l be the histograms of X and Y at resolution l. $H_X^l(i)$ and $H_Y^l(i)$ are the numbers of points from X and Y that fall into the ith cell of the grid. The amount of matches at level l is given by the *histogram intersection function*[8]:

$$\Gamma(H_X^l, H_Y^l) = \sum_{i=1}^{D} min(H_X^l(i), H_Y^l(i)) \tag{1}$$

After counting the amount of matches that occurs at each level of the pyramid, these amounts are summed using weights for each level. The weight associated to level l is set to $\frac{1}{2_{L-l}}$. The intention is penalizing matches found in larger cells because they involve increasingly dissimilar features [8], in other words, it weights features at higher levels more highly, reflecting the fact that higher levels localize the features more precisely Then, with the penalization, the *pyramid match kernel* is defined as:

$$K^L(X,Y) = \Gamma^L + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}} (\Gamma^l - \Gamma^{l-1}) \tag{2}$$

$$= \frac{1}{2^L}\Gamma^0 + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}} \Gamma^l \tag{3}$$

Finally, all image feature pyramids are concatenated and utilized as input of the classifier. The low-level features used are the SIFT features. In our approach, the histogram intersection function (Eq. 1) is not used as the *pyramid match kernel*. Instead, we use a linear SVM kernel.

3 Experiments

To evaluate the proposed method, we use the ASL Finger Spelling Dataset [9]. This data set contains 500 samples for each of 24 signs, recorded from five 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.

Due to the variety in the orientation when the signal is performed, signs became strongly similar. Fig. 1 shows some examples and there is possible to see the variety in size, background and orientation, the most similar signs *a*, *e*, *m*, *n*, *s* and *t*. The examples are taken from the same user. It is easy to identify similarities 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. 1. 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.

To be able to compare the results achieved in this work with previously published approaches, we conduct experiments considering three evaluation protocols. The first uses 10% of the dataset for training, the second considers 50% of the dataset for training. Finally, the third protocol uses a 5-fold cross validation. For all experiments, we defined the following specifications:

- We use a dense SIFT descriptor to extract features in intensity images, they were computed over a grid of 16×16 patches with spacing of 8 pixels, as defined in [8].
- We use the k-means algorithm for clustering a random subset of patches to form a visual vocabulary, this was constructed using a 5% of data set. We test with different vocabulary sizes: 10, 30, 50, 80, 100 and 150.
- A three level pyramid ($L = 2$) was used to construct the multi-resolution weighted and non-weighted histograms.

– In the classification stage, we use a linear kernel. We also use different percentages of samples for training and testing. The library LIBSVM (a library for Support Vector Machines)] [12] was used in our implementation.

Using the aforementioned evaluation protocols and the parameter setting, we study aspects such as the influence of the weight associate to each level of the pyramid match kernel and the number of visual words used to build the dictionary. Finally, we compare the results achieved by the proposed method to others state-of-the-art approaches.

Weights in the pyramid match kernel. According to the results obtained for all evaluation protocols, shown in Tables 1, 2 and 3, the influence of using the weights associated with the level of the spatial pyramid [8] is very small for the finger spelling recognition. We believe that such results are due to the low complexity of the images since most of them do not have a complex background, which reduces the need for penalizing large cells. Therefore, we will not consider the weights in the remaining experiments.

Number of visual words. Different from the weights in the pyramid match kernel, the number of codewords used in the visual dictionary presents a large influence in the accuracy, as can be seen in Tables 1, 2 and 3. The results indicate that the more codewords the better the results are. Even though the results were still improving when 150 codewords were considered, we stop increasing the number since the results were not improving much and the computation cost was increasing significantly. Therefore, we will use 150 codewords to compare our results to other previously published methods.

Table 1. Accuracy (Acc.) and standard deviation (Std. Dev.) of the classification using intensity information with different dictionary size using 10% of training set.

Number	Weight		No weight	
Words	Acc. (%)	Std. Dev.	Acc. (%)	Std. Dev.
10	75.6	0.09	75.7	0.09
30	85.9	0.07	85.7	0.07
50	89.4	0.06	88.9	0.06
80	90.9	0.05	91.00	0.04
100	91.4	0.05	91.5	0.05
150	92.3	0.04	92.5	0.04

Comparisons to State-of-the-art Approaches. We can see in Table 4 the comparison of our approach in three different scenarios (protocols defined earlier). In the first, a 10% of dataset is used for training our SVM classifier. Our approach obtained the highest accuracy, outperforming the methods based on intensity and depth information proposed in [7, 10, 13]. In the second experiment, a 50% of the data is used as training set, again our approach outperforms [5]

Table 2. Accuracy (Acc.) and standard deviation (Std. Dev.) of the classification using intensity information with different dictionary size using 50% of training set.

Number	Weight		No weight	
Words	Acc. (%)	Std. Dev.	Acc. (%)	Std. Dev.
10	81.5	0.08	81.5	0.08
30	92.2	0.05	92.0	0.04
50	94.8	0.04	94.4	0.04
80	95.9	0.03	95.9	0.03
100	96.5	0.03	96.5	0.03
150	97.1	0.02	97.2	0.02

Table 3. Accuracy (Acc.) and standard deviation (Std. Dev.) of the classification using intensity information with different dictionary size with 5-fold cross-validation.

Number	Weight		No weight	
Words	Acc. (%)	Std. Dev.	Acc. (%)	Std. Dev.
10	82.5	0.07	82.5	0.07
30	93.2	0.04	93.0	0.04
50	95.6	0.03	95.4	0.03
80	96.8	0.03	96.8	0.03
100	97.2	0.02	97.3	0.02
150	97.8	0.02	97.9	0.02

and [7]. Finally, in the third experiment, a 5-fold cross-validation was performed. Our approach outperforms in more than 10% the results of [14] using RGB information.

We believe that the proposed method was able to achieved the best results compared to the state-of-the-art methods due to the capability of capturing multi-scale information through the spatial pyramid. This way, both fine and coarse information can be encoded in the codewords simultaneously.

4 Conclusions and Future Works

In this paper, we proposed a method for Finger Spelling Recognition from intensity information using the spatial pyramid matching descriptor. A spatial pyramid is a collection of orderless feature histograms computed over cells defined by a multi-level recursive image decomposition that are classified by a SVM.

The experiments have shown that the proposed approach is promising for sign language based only on intensity information. In all experiments, we achieved the highest recognition rate compared with methods that combine intensity and depth information. Based on the experiments, we can say that the influence of the

Table 4. Accuracy (Acc.) and standard deviation (Std. Dev.) of the classification using different protocols. A: 10% of the data set used for training; B: 50% of the data set used for training; C: 5-fold cross-validation.

Protocol	Method	Acc. (%)	Std. Dev.
A	Depth [7]	75.60	0.26
	RGB [7]	79.08	0.25
	RGB-D [7]	88.54	0.17
	Zhu & Wong [10]	88.90	0.39
	Estrela et al. [13]	71.51	-
	proposed approach	**92.50**	**0.04**
B	Pugeault & Bowden[5]	75.00	-
	Depth [7]	86.85	0.17
	RGB [7]	91.58	0.16
	RGB-D [7]	96.77	0.09
	proposed approach	**97.1**	**0.02**
C	Depth [14]	62.70	0.47
	RGB [14]	85.18	0.16
	RGB-D [14]	91.26	0.18
	proposed approach	**97.9**	**0.02**

weight associated to spatial pyramid is very small for finger spelling recognition. This probably occurs due to the low complexity of the images, most of them do not have a complex background. The results also show that the more codewords the better the results are. Although the results were still improving, we choose a 150 codewords dictionary because the results were not improving too much while the computational cost increases.

As a future work, we intend to extend our recognition methods to work with dynamic scenes in which the ambiguities with some characters such as J and Z can be solved.

Acknowledgments. The authors are thankful to CNPq, CAPES and FAPEMIG (Grants APQ-02292-12 and APQ-01294-12), Brazilian funding agencies and to the Federal University of Ouro Preto for supporting this work.

References

1. J. Isaacs and S. Foo, "Hand pose estimation for american sign language recognition," *36th Southeastern Symposium on System Theory*, pp. 132–136, 2004.
2. D. Dahmani and S. Larabi, "User-independent system for sign language finger spelling recognition," *Journal of Visual Communication and Image Representation*, no. 0, pp. –, 2014.
3. S. Auephanwiriyakul, S. Phitakwinai, W. Suttapak, P. Chanda, and N. Theera-Umpon, "Thai sign language translation using scale invariant feature transform

and hidden markov models," *Pattern Recognition Letters*, vol. 34, no. 11, pp. 1291 – 1298, 2013.

4. 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)*, pp. 383–390, 2011.
5. 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)*, pp. 1114–1119, IEEE, 2011.
6. 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)*, WACV '11, (Washington, DC, USA), pp. 66–72, IEEE Computer Society, 2011.
7. K. Otiniano-Rodríguez and G. Cámara-Chávez, "A robust kernel descriptor for finger spelling recognition based on rgb-d information," *International Journal of Computer Science & Information Security*, vol. 11, pp. 1–7, December 2013.
8. S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, (Washington, DC, USA), pp. 2169–2178, IEEE Computer Society, 2006.
9. R. B. Nicolas Pugeault, "ASL finger spelling dataset." `http://personal.ee.surrey.ac.uk/Personal/N.Pugeault/index.php`, last visit: April 29, 2013.
10. X. Zhu and K.-Y. K. Wong, "Single-frame hand gesture recognition using color and depth kernel descriptors," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR)*, pp. 2989–2992, IEEE, 2012.
11. K. Grauman and T. Darrell, "Pyramid match kernels: Discriminative classification with sets of image features," in *ICCV*, 2005.
12. 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`.
13. B. N. Estrela, G. Cmara-Chvez, M. F. M. Campos, W. R. Schwartz, and E. R. Nascimento, "Sign language recognition using partial least squares and rgbd information," in *Proceedings of the IX Workshop de Viso Computacional (WVC2013)*, 2013.
14. K. Otiniano-Rodríguez and G. Cámara-Chávez, "Finger spelling recognition from RGB-D information using kernel descriptor," in *Proceedings of the SIBGRAPI 2013 (XXVI Conference on Graphics, Patterns and Images)*, pp. 1–7, 2013.