Multimodal Learning for Sign Language Recognition

Pedro M. Ferreira(✉), Jaime S. Cardoso, and Ana Rebelo

INESC TEC, Porto, Portugal
{pmmf,jaime.cardoso,arebelo}@inesctec.pt

Abstract. Sign Language Recognition (SLR) has becoming one of the most important research areas in the field of human computer interaction. SLR systems are meant to automatically translate sign language into text or speech, in order to reduce the communicational gap between deaf and hearing people. The aim of this paper is to exploit multimodal learning techniques for an accurate SLR, making use of data provided by Kinect and Leap Motion. In this regard, single-modality approaches as well as different multimodal methods, mainly based on convolutional neural networks, are proposed. Experimental results demonstrate that multimodal learning yields an overall improvement in the sign recognition performance.

Keywords: Sign Language Recognition · Multimodal learning · Convolutional neural networks · Kinect · Leap Motion

1 Introduction

Sign language (SL) is an integral form of communication especially used by hearing impaired people within deaf communities worldwide. It is a visual means of communication, with its own lexicon and grammar, that combines articulated hand gestures along with facial expressions to convey meaning. As most of hearing people are unfamiliar with SL, deaf people find it difficult to interact with the hearing majority. In this regard, Sign Language Recognition (SLR) has becoming an appealing topic in modern societies. Its main purpose is to automatically translate the signs from video or images into the corresponding text or speech. This is important not only to bridge the communicational gap between deaf and hearing people but also to increase the amount of contents to which the deaf can access (e.g., educational tools or games for deaf and visual dictionaries of SL).

The SLR task can be addressed by using wearable devices or vision-based approaches. Vision-based SLR is less invasive since there is no need to wear cumbersome devices that might affect the natural signing movement. A vision-based SLR system is typically composed by three main building blocks: (i) hand segmentation and/or tracking, (ii) feature extraction, and (iii) sign recognition. The SLR problem was first addressed by the computer vision community by means of just using the colour information of images and videos [1,2].

L.A. Alexandre et al. (Eds.): IbPRIA 2017, LNCS 10255, pp. 313–321, 2017.
DOI: 10.1007/978-3-319-58838-4_35

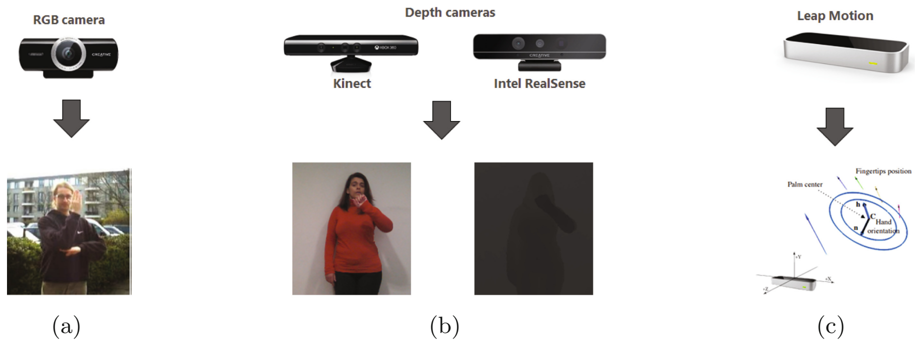


Fig. 1. Vision-based SLR systems: (a) colour information provided by RGB cameras, (b) colour and depth information provided by depth cameras, and (c) hand position and orientation provided by Leap Motion.

More recently, the emergence of low-cost consumer depth cameras (e.g., Microsoft's Kinect) has promoted the development of several approaches that try to combine colour and depth information (see Fig. 1). Bergh and Gool [3] demonstrated that depth information can be used together with colour information to increase the recognition accuracy, especially when there is superposition between hands and face. In [4], multiple depth-based descriptors are fed into a SVM classifier for gesture recognition.

The recent introduction of the Leap Motion has launched new research lines for gesture recognition. Instead of a complete depth map, the Leap Motion sensor directly provides the 3D spatial positions of the fingertips and the hand orientation with quite accuracy ($\approx 200\,\mu\text{m}$) (see Fig. 1). One of the first studies referring to the utilization of Leap Motion for SLR has been presented in [5]. The authors stated that, although Leap Motion may have a great potential for sign recognition, it is not always able to recognize all fingers in some hand configurations. In order to overcome that limitation, Marin *et al.* [6,7] combined the input data from Leap Motion with Kinect.

In this work, we extent the ideas proposed in [6,7], improving their results. In particular, our main contributions are:

- We explore the concept of convolutional neural networks (CNNs) for recognizing SL, in two different ways. First, CNNs are used to directly classify the sign. Second, CNNs are used as feature extractor, avoiding the hand-craft feature extraction process and the inherent difficulty of designing reliable features to the large variations of hand gestures.
- We develop a multimodal learning framework for the SLR problem, making use of data provided by both Kinect (colour + depth) and Leap Motion.
- We performed a comparative study between single-modality and multimodal learning techniques, in order to demonstrate the effectiveness of multimodal learning in the overall sign recognition performance.

The paper is organized in four sections including the Introduction (Sect. 1). In Sect. 2, the proposed SLR methods are fully described. Section 3 reports the experimental results. Finally, conclusions and some topics for future work are presented in Sect. 4.

2 Methodology

The aim of this paper is to explore the potential of multimodal learning for SLR. To accomplish this purpose, single-modality approaches as well as different multimodal methods, to fuse them at different levels, are proposed. Multimodal techniques include data-level, feature-level and decision-level fusion methods.

2.1 Single-Modality Sign Recognition

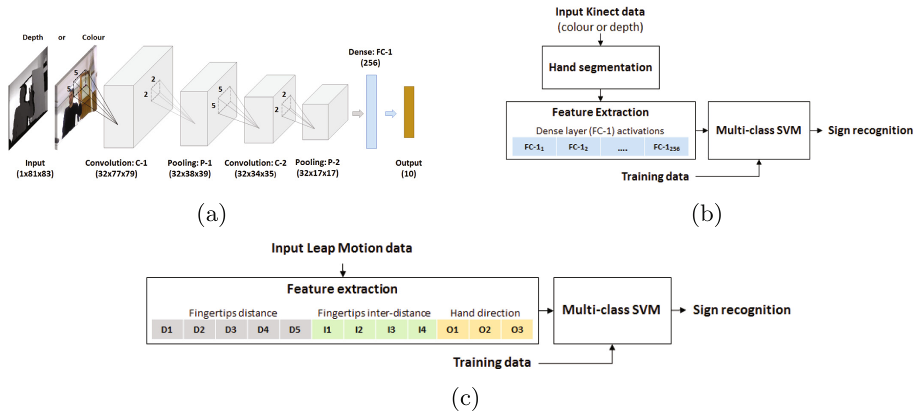
2.1.1 Kinect Modalities (Colour and Depth)

In this work, convolutional neural networks (CNN) were explored in two different ways. In the first approach, a CNN is used to directly classify the sign. In the second approach, the CNN is used as a feature extractor.

Both Kinect modalities, colour and depth, require a pre-processing step in order to segment the hands, from the noisy background of the image, before feature extraction and sign recognition. In the first step, a skin colour model is used to distinguish skin pixels from background pixels. This skin colour binarization is used to filter the depth map. Then, the hand segmentation is performed on the filtered depth map by just using depth information.

CNN Model as Classifier. The implemented neural network follows the traditional CNN architecture for classification, typically starting from several sequences of convolution-pooling layers to fully connected layers [8]. Hence, the implemented CNN is composed by two convolution layers and one fully connected layer (or dense layer), in which each convolution layer is followed by a 2×2 max-pooling layer. Both convolution layers have the same filters' number and size. Finally, the last layer of the CNN is a softmax output layer. The output layer contains the output probabilities for each class label. The output node that produces the largest probability is chosen as the overall classification. The architecture of implemented CNN is illustrated in Fig. 2a. During the training stage, several regularization techniques, such as the L2 norm, data augmentation and dropout [9], were applied to prevent overfitting.

CNN Model as Feature Descriptor. The later layers of a CNN seem to learn visually semantic attributes of the input [8]. Hence, these intermediate representations can be used as a generic feature descriptor. Many research works [8] stated that these CNN features are better than hand-crafted features, such as SIFT or HoG, for several computer vision tasks. In here, the CNN is used as feature extractor instead of being used as a classifier. More concretely, the



(a)

(b)

(c)

Fig. 2. Single-modality sign recognition: (a) CNN model as classifier; (b) CNN model as feature descriptor (methodologies applied to both Kinect modalities); and (c) Leap Motion sign recognition methodology.

activations of the last dense layer (FC-1) are extracted to be used as a feature descriptor (see Fig. 2a). For sign recognition, this CNN feature descriptor is fed into a multi-class SVM classifier (see Fig. 2b).

2.1.2 Leap Motion

Unlike Kinect, Leap Motion does not provide a complete depth map, instead it directly provides a set of relevant features of hand and fingertips. In this paper, 3 different types of features computed from the Leap Motion data are used:

1. **Fingertip distances** $D_i = \|F_i - C\|, i = 1, ..., N$; where N denotes the number of detected fingers and D_i represents the 3D distances between each fingertip F_i and the hand centre C.
2. **Fingertip inter-distances** $I_i = \|F_i - F_{i+1}\|, i = 1, ..., N-1$; represent the 3D distances between consecutive fingertips.
3. **Hand direction** O: represents the direction from the palm position toward the fingers. The direction is expressed as a unit vector pointing in the same direction as the directed line from the palm position to the fingers.

Both distance features are normalized by signer (user), according to the maximum fingertip distance and fingertip inter-distance of each user. This normalization is performed to make those features robust to people with different hand's size. Then, these 3 features are used as input into a multi-class SVM classifier for sign recognition. The block diagram of the implemented Leap Motion-based sign recognition approach is illustrated in Fig. 2c.

2.2 Multimodal Sign Recognition

The data provided by Kinect and Leap Motion have quite complementary characteristics, since while Leap Motion provides few accurate and relevant keypoints, Kinect produces both a colour image and a complete depth map with a large number of less accurate 3D points. Therefore, we intend to exploit them together for SLR purposes.

According to the level of fusion, multimodal fusion techniques can be roughly grouped into three main categories: (i) data-level, (ii) feature-level, and (iii) decision-level fusion techniques [10]. As described in the following, we propose multimodal approaches of each fusion category for the SLR task, making use of 3 modalities (i.e. colour, depth and Leap Motion data).

2.2.1 Data-Level Fusion

The purpose of data-level fusion is to merge data from different modalities at an early stage. As illustrated in Fig. 3a, this methodology simply consists in the concatenation of the RGB colour image with the depth map, which results in a 4-dimensional matrix. In this approach, just both Kinect modalities (i.e. colour and depth) are considered for fusion, since the data dimensions of Leap Motion are incompatible.

2.2.2 Feature-Level Fusion

In general, feature-level fusion is characterized by three phases: (i) learning a representation, (ii) supervised training, and (iii) testing [10]. According to the order in which phases (i) and (ii) are made, feature-level fusion techniques can be roughly divided into two main groups: (1) End-to-end fusion, where the representation and the classifier are learned in parallel - see Fig. 3b; and (2) Multi-step fusion, where the representation is first learned and then the classifier is learned from it - see Fig. 3c.

End-to-End Fusion. The underlying idea of this approach is to learn an end-to-end deep neural network. In our scenario, the neural network has multiple input-specific pipes (one for each data type: colour, depth and Leap Motion), in which each input type is processed by its specific neural net. While colour and depth are both processed by a CNN, the Leap Motion data is processed by a classical neural net with one hidden layer. Then, the last hidden layers of each pipe are concatenated followed by one additional fully connected layer. All the layers are trained together end-to-end. The architecture of the implemented neural network is represented in Fig. 3b.

Multi-step Fusion. As in the end-to-end approach, a shared (multimodal) representation vector is created, by concatenating the last hidden layers of each model previous trained individually. Then, for sign recognition, the multimodal representation vector is fed into an additional classifier (i.e. a multi-class SVM). The multi-step feature-level fusion scheme is depicted in Fig. 3c.

2.2.3 Decision-Level Fusion

The purpose of decision-level fusion is to learn a specific classifier for each modality and, then, to find a decision rule between them. In this paper, we apply this

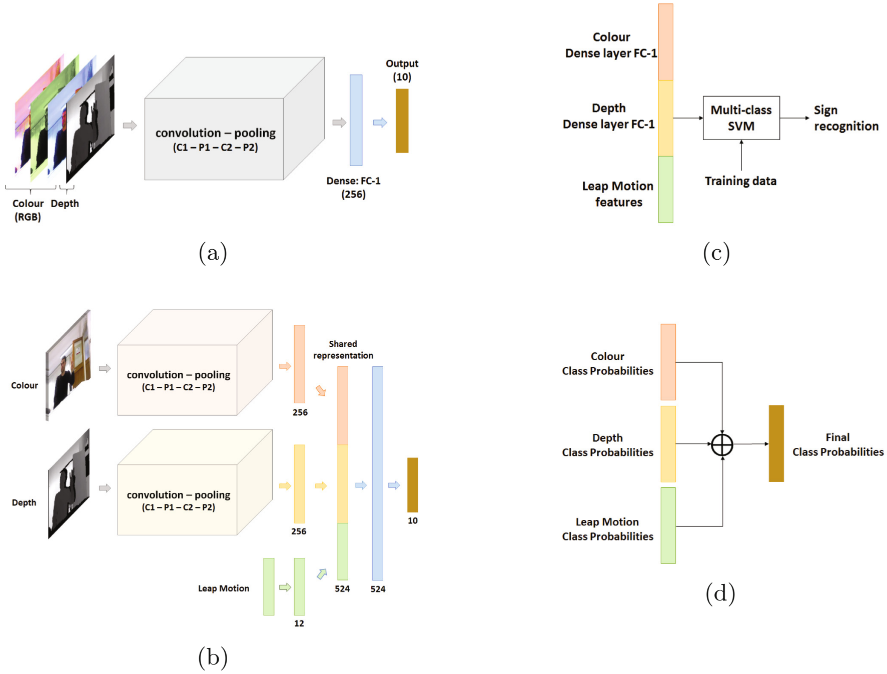

Colour (RGB)
Depth
convolution – pooling (C1 – P1 – C2 – P2)
Dense: FC-1 (256)
Output (10)
(a)
Colour
Depth
convolution – pooling (C1 – P1 – C2 – P2)
convolution – pooling (C1 – P1 – C2 – P2)
256
256
Shared representation
Leap Motion
12
524
524
10
(b)
Colour Dense layer FC-1
Depth Dense layer FC-1
Leap Motion features
Multi-class SVM
Training data
Sign recognition
(c)
Colour Class Probabilities
Depth Class Probabilities
Leap Motion Class Probabilities
Final Class Probabilities
(d)

Fig. 3. Multimodal sign recognition: (a) Data-level fusion; (b) Decision-level fusion, where ⊕ is an aggregate operator representing the decision rule for fusion; (c) End-to-end feature-level fusion; and (d) Multi-step feature-level fusion.

concept making use of the output class probabilities of the models designed individually for each modality under analysis. Then, two main kinds of decision rules, to combine these class probabilities, were implemented: (1) pre-defined decisions rules, and (2) decision rules learned from the data (see Fig. 3d).

Pre-defined Decision Rules. Herein, two different pre-defined decision rules were implemented. In the first approach, the final prediction is given by the argument that maximizes the averaged class probabilities. In the second approach, the final prediction is given by the model with the maximum confidence. The confidence of a model in making a prediction is measured by its highest class probability.

Learned Decision Rule. The underlying idea of this approach is to learn a decision rule from the data. Therefore, a descriptor that concatenates the class probabilities, extracted from the individual models of each modality, is created and, then, used as input into a multiclass SVM classifier for sign recognition.

3 Experimental Results

The experimental evaluation of the proposed methodologies was performed in a public Microsoft Kinect and Leap Motion hand gesture recognition database [6,7]. The database is composed by 10 static gestures from the ASL. Each sign was performed by 14 different people, and repeated 10 times, which results in a total of 1400 gestures. In order to ensure signer independence, the dataset is divided into a training set of 1000 images from 10 people, and a test set of 400 images from the other 4 people. The training set is further divided in half, resulting in two subsets: one for training all single-modality methods and another for training the multimodal techniques that require input from single-modality methods, such as the feature-level and decision-level fusion approaches.

The implementation of the deep neural networks is based on Theano. The Nesterov's Accelerated Gradient Descent with momentum is used for optimization, and the categorical cross-entropy is used as the loss function. The adopted SVM classifier consists in a multi-class SVM classifier based on the one-against-one approach, in which a nonlinear Gaussian Radial Basis Function (RBF) kernel is used. The parameters (C, γ) of the RBF kernel are estimated by means of a grid search approach and cross-validation on the training set.

3.1 The Potential of Multimodal Learning

In order to access the potential of multimodal learning in the SLR context, we computed the rate of test signs for which each single-modality method made a correct prediction while the others were wrong. As presented in Table 1a, these results clearly demonstrate that there is a relative big potential to tackle the SLR problem via multi-modality. In particular, there is a higher complementarity between each Kinect modality (i.e. colour or depth) with the Leap Motion rather than between both Kinect modalities. For instance, there are 4.25% and 5.75% of test instances for which Leap Motion made correct predictions while colour and depth made incorrect ones, respectively.

3.2 Discussion

The experimental results of the proposed single-modality and multimodal sign recognition methodologies are presented in Table 1b and c, respectively. The results are reported in terms of classification accuracy (Acc), which is given by the ratio between the number of correctly classified signs t and the total number of test signs n: $Acc\% = \frac{t}{n} \times 100$. A first observation, regarding single-modality approaches, is that both colour and depth outperform Leap Motion, with accuracies of 94.75%, 91.75% and 82.00%, respectively. However, it should be noticed that Leap Motion sign recognition does not require any kind of preprocessing in order to segment the hand from the background for feature extraction. The most interesting observation is that multimodal fusion often promotes an overall improvement in the sign recognition accuracy. These results clearly demonstrate the complementarity between the three modalities. Typically, the classification

Table 1. Experimental assessment of the proposed recognition methods. (a) The potential of multimodal learning, expressed by the rate of test instances for which modality B made correct predictions while modality A made incorrect ones. (b) and (c) Experimental results of the proposed single-modality and multimodal recognition approaches, respectively. The results are presented in terms of classification accuracy (%).

(a)

Modality A	Modality B	Multi-modality potential (%)
Colour	Depth	3.00
Colour	Leap Motion	4.25
Depth	Colour	4.75
Depth	Leap Motion	5.75
Leap Motion	Colour	18.5
Leap Motion	Depth	18.25

(b)

Modality	Method	Acc (%)
Colour	CNN C†	93.50
	CNN FEAT‡	**94.75**
Depth	CNN C	**91.75**
	CNN FEAT	90.75
Leap Motion	-	**82.00**

† CNN as classifier.

‡ CNN as feature extractor.

(c)

Proposed multimodal learning methodologies			
Fusion-level	**Method**	**Involved Modalities**	**Acc (%)**
Data	-	C + D	89.75
Feature	**End-to-end**	C + D	93.00
		C + D + L	94.25
	Multi-step	C + D	96.25
		C + D + L	96.75
Decision	**Average rule**	C + D	96.00
		C + D + L	**97.00**
	Highest confidence	C + D	96.00
		C + D + L	96.50
	Learned decision rule	C + D	96.25
		C + D + L	96.75
State-of-the-art methodologies			
	Marin *et al.* 2014 [6]		91.28
	Marin *et al.* 2015 [7]		96.50

accuracy increases as each modality is added to the recognition scheme. In particular, the decision-level fusion scheme, with the average decision rule, provides the best overall classification accuracy ($Acc = 97.00\%$). Still, regarding multimodal fusion techniques, it is possible to observe that, in general, decision-level fusion performs better than data-level and feature-level fusion. In fact, data-level fusion resulted in a worst model than the best single-modality method, with an Acc of 89.75%. These worst results are probably due to the curse of dimensionality, as the dimension of the input features in this model is considerable higher than in the others. Likewise, the end-to-end feature-level fusion approach also performed worst than the best single-modality method. This result might seem quite unexpected; however, a multimodal neural net architecture with multiple input-specific pipes has potentially more local minima which may explain the unsatisfying results. The initialization of the input specific weights from pre-trained single-modality networks might improve the results. Finally, it is important to stress that the best implemented multimodal fusion approach outperformed both state-of-art methods [6,7], with an Acc of 97.00% against 91.28% and 96.50%, respectively.

4 Conclusions

This paper addresses the topic of static SLR, by exploring multimodal learning techniques, making use of data from 3 distinct modalities: (i) colour; (ii) depth, both from Kinect; and (iii) Leap Motion data. In this regard, single-modality approaches as well as different multimodal methods, to fuse them at different levels, are proposed. Multimodal techniques include data-level, feature-level and decision-level fusion techniques. Experimental results suggest that both Kinect modalities are more discriminative than the Leap Motion data. However, the most interesting observation is that, in general, multimodal learning techniques outperform single-modality methods. In particular, the proposed decision-level fusion scheme, with the average decision rule, achieved the best results ($Acc = 97.00\%$) and outperforms the current state-of-the-art methods. As future work, it is expected to extend the proposed methodologies for dynamic signs.

Acknowledgements. This work was funded by the Project “NanoSTIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE-01-0145-FEDER-000016” financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF), and also by Fundação para a Ciência e a Tecnologia (FCT) within PhD and BPD grants with numbers SFRH/BD/102177/2014 and SFRH/BPD/101439/2014.

References

1. Cooper, H., Bowden, R.: Large lexicon detection of sign language. In: Lew, M., Sebe, N., Huang, T.S., Bakker, E.M. (eds.) HCI 2007. LNCS, vol. 4796, pp. 88–97. Springer, Heidelberg (2007). doi:10.1007/978-3-540-75773-3_10
2. Adithya, V., Vinod, P.R., Gopalakrishnan, U.: Artificial neural network based method for Indian sign language recognition. In: 2013 IEEE Conference on Information Communication Technologies (ICT), pp. 1080–1085 (2013)
3. den Bergh, M.V., Gool, L.V.: Combining RGB and ToF cameras for real-time 3D hand gesture interaction. In: 2011 IEEE Workshop on Applications of Computer Vision (WACV), pp. 66–72, January 2011
4. Dominio, F., Donadeo, M., Zanuttigh, P.: Combining multiple depth-based descriptors for hand gesture recognition. Pattern Recog. Lett. **50**, 101–111 (2014). Depth Image Analysis
5. Potter, L.E., Araullo, J., Carter, L.: The leap motion controller: a view on sign language. In: Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration, OzCHI 2013, pp. 175–178. ACM, New York (2013)
6. Marin, G., Dominio, F., Zanuttigh, P.: Hand gesture recognition with leap motion and kinect devices. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 1565–1569, October 2014
7. Marin, G., et al.: Hand gesture recognition with jointly calibrated leap motion and depth sensor. Multimedia Tools Appl. **75**(22), 14991–15015 (2015)
8. Srinivas, S., Sarvadevabhatla, R.K., Mopuri, K.R., Prabhu, N., Kruthiventi, S., Radhakrishnan, V.B.: A taxonomy of deep convolutional neural nets for computer vision. Front. Robot. AI **2**(36), 1–13 (2016)
9. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**, 1929–1958 (2014)
10. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: International Conference on Machine Learning, vol. 6 (2011)