MORA: A Generative Approach to Extract Spatiotemporal Information Applied to Gesture Recognition

Igor L. O. Bastos, Victor H. C. Melo, Gabriel Resende Gonçalves, William Robson Schwartz
Smart Surveillance Interest Group, Smart Sense Laboratory, Department of Computer Science
Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
Email: {igorlobastos, victorhcmelo, gabrielrg, william}@dcc.ufmg.br

Abstract

Gestures are related to a non-verbal language used on the interaction between subjects. Due to its applicability in several contexts, gesture recognition has been investigated by different researches, often investing on the capture of motion and appearance on videos. However, most of these methods do not properly explore the well-defined gesture temporal structure and are not suitable to deal with an increasing number of classes. Thus, we propose the Multi-Output Recurrent Autoencoders (MORA), an approach that relies on the representation of each gesture class independently. MORA employs a specific autoencoder model per class, composed by convolutional (3D) and a Gated Recurrent Unit (GRU) layer, what allows spatiotemporal information extraction and scalability in terms of number of classes. To validate MORA, experiments are conducted on SKIG and ChaLearn IsoGD datasets, for which the approach achieved accuracies comparable to state-of-the-art methods.

1. Introduction

Gesture recognition corresponds to a mathematical interpretation of a human motion by a computer device, involving hands, arms, face, head and/or body [11]. Gestures comprise both temporal and spatial information denoted by changes in appearance and motion over time, characterized by a structured time disposition, where the order of events (sub-actions) is relevant to determine their labels [12].

Being able to recognize gestures allows a wide range of applications in different contexts, such as navigation on virtual environments, development of aid systems for hearing impaired, sign language recognition, surveillance monitoring and biometric validation [14, 21, 23]. For this reason, gesture recognition has been investigated in a wide range of approaches, which vary in terms of features and learning algorithms employed to perform the task [12, 1, 13, 3].

Despite major advances achieved [15], developing an universal model to recognize gestures is a difficult task due to problems such as illumination and acquisition conditions, inconsistent behavior among users, cultural gesture specificities, and large vocabularies [22]. In addition, assembling gesture recognition datasets is an expensive task and most of them do not comprise many gestures [16]. As a consequence, approaches usually do not invest on scalability and require to be completely retrained to handle new gestures, leading to methods that are not suitable, for instance, for sign language recognition, since these languages are extremely changeable and adhere to cultural specificities from people and places where they are employed [17].

Focusing on the aforementioned scalability issue, the present work introduces *Multi-Output Recurrent Autoencoders* (MORA)[1]. The approach consists on the application of an autoencoder model to characterize each gesture class in an independent and scalable way, in the sense that increasing the number of classes only requires the training of new autoencoders without any impact on already trained models. To properly gather temporal/spatial information, these models contain stacked 3D convolutional layers. On the mid-level representation of the model, a Gated Recurrent Unit (GRU) [4] layer is employed to exploit the well-structured time behavior of gestures. Since different aspects are relevant for the recognition of a gesture, MORA captures appearance, motion and depth through the employment of a multi-output loss function, which considers the reconstruction of RGB, optical flow and depth videos.

To validate our approach, tests are conducted on the SKIG [9] and ChaLearn IsoGD [20] datasets, for which the approach achieves 97.20% and 66.16% of accuracy, respectively, being comparable to the-state-of-art methods, with the advantage of being scalable.

2. Related Work

Recent advances on machine vision technology and the high applicability of gesture recognition have stimulated

[1] Code is available: https://github.com/igorcrexito/MORA

the development of methods tackling this task in the video domain. Among them, hand-crafted spatiotemporal features [9, 5] and deep learning-based methods [12, 22, 13] have been proposed. Most of the approaches focus on the extraction/learning of shape, appearance and motion cues to determine the label of a gesture [12].

The accurate outcomes of deep learning approaches regarding gesture recognition in videos have stimulated the development of increasingly complex and effective models, which usually apply spatiotemporal operations to learn features that better distinguish dataset classes [12]. An important point on the assembling of these models lies on the composition of the extraction/learning algorithm architecture. For instance, Tranand et al. [2] proposed a straightforward architecture based on stacked 3D convolutions, achieving accurate results in different video analysis and gesture recognition tasks.

On the other hand, many approaches invest on the exploitment of the strong temporal correlation between subevents in gesture videos, leading to recurrent models that have been proposed and achieved state-of-the-art results for most gesture recognition datasets. Molchanov et al. [12] proposed a model that extracts spatiotemporal features from video clips and propagate such information through a recurrent layer followed by a softmax. Although simplistic, this architecture proved to be effective for recognition and detection of gestures on datasets such as SKIG [9], ChaLearn [20] and Multimodal Dynamic Gesture [12].

Aiming at capturing the temporal correlation in gesture videos, Nishida and Nakayama [13] investigated an architecture composed by LSTM layers to handle videos with variable-length gestures. To create a spatiotemporal representation, multiple temporal modalities are fused, which led to a high accuracy on the SKIG dataset [9]. Similarly, Zhang et al. [22] proposed an architecture to learn spatiotemporal features based on the application of purely spatial convolutions and bidirectional convolutional LSTM layers to explore the well-defined time structure of gesture videos. This model was tested with several fusion modalities and achieved state-of-the-art results for the SKIG [9] and the ChaLearn IsoGD [20] datasets.

Despite similarities with previously proposed recurrent models, our approach differs from them due to the application of non-discriminative models and a multi-output loss function that considers the reconstruction of appearance, motion and depth, relevant aspects for gesture recognition. Therefore, the recognition of a gesture is based on the reconstruction error obtained for each dataset class. In addition, with the employment of a specific autoencoder model for each dataset class, MORA stands out when compared to other methods in terms of scalability, once it does not require the retraining of previously learned models when new classes (i.e., new gestures) are considered.

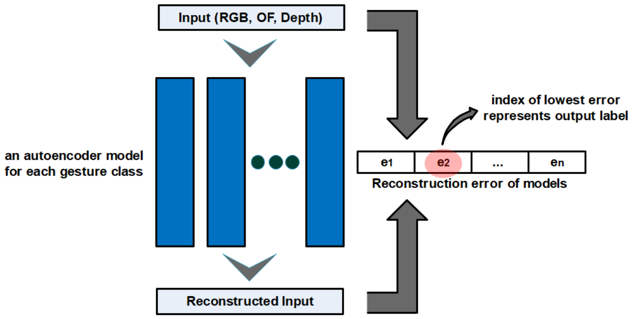
Input (RGB, OF, Depth)
an autoencoder model for each gesture class
index of lowest error represents output label
e1
e2
...
en
Reconstruction error of models
Reconstructed Input

Figure 1: Proposed approach. The model with the lowest reconstruction error indicates the class label of an input video.

3. Proposed Approach

The present research proposes a novel strategy to recognize gestures based on the employment of multiple autoencoder models. For each class (gesture), a model is assembled and trained. Since appearance, motion and depth play a relevant role on gesture recognition, they are used as inputs for our models, as well as outputs to be reconstructed. On test phase, the reconstruction error is used for classification, being computed with the sum of differences between each input and its reconstructed output. The index of the model with the lowest reconstruction error represents the label of the video, as illustrated in Figure 1.

The first step of the present approach corresponds to the processing of input data. Most gesture recognition databases, such as SKIG and ChaLearn, provide RGB and depth videos. To better capture motion of gestures, Farneback optical flow [7] is computed over RGB videos. Thus, our final input consists of videos representing these modalities: RGB, Optical Flow and Depth. For each one, video frames are resized to $112 \times 112 \times 3$ (rows, columns and channels). To avoid redundant information, an uniform subsampling is performed, producing 32-frame videos.

Gestures comprise a well-defined time structure which must be properly captured for accurate recognition. With this goal in mind, we assemble our models with a recurrent GRU layer to propagate information through each timestep (shown in Figure 2). Granted that, we adjust the input of our models accordingly to process 8-frame clips each time, resulting in four timesteps for each video, as they are sampled with 32 frames. This enables the autoencoder model, for each class of gesture, to learn spatiotemporal information from the three inputs, i.e., RGB, optical flow and depth, and reconstruct them at the end. This reconstruction minimizes a multi-output loss function, which is based on the sum of the mean squared error between each input and their reconstructed counterpart. It is important to emphasize that both the state and the output of the GRU layer are used to produce a combined representation for each timestep, resulting in a vector with 784 dimensions.

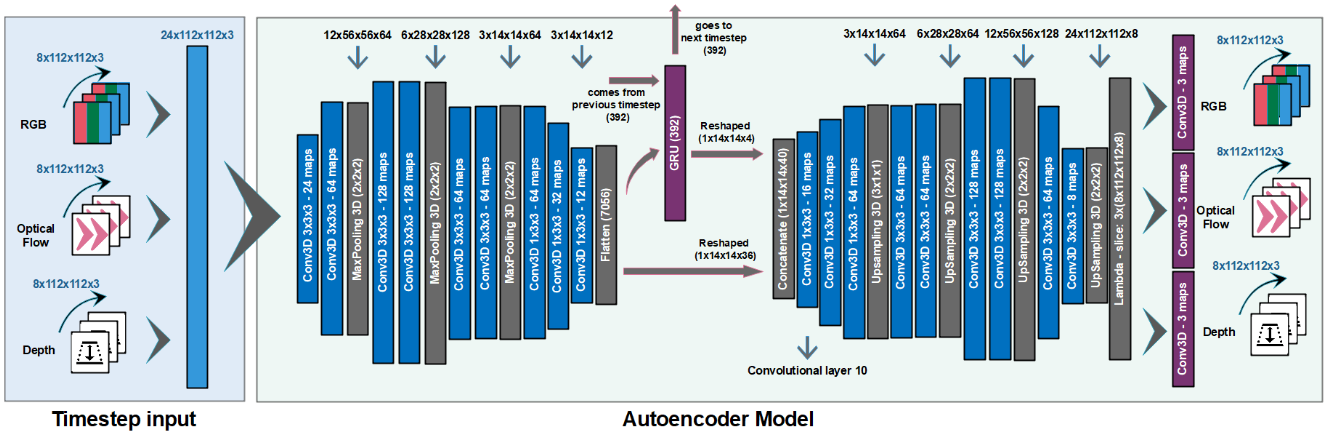

Figure 2: Proposed autoencoder. Layers employ ReLU (blue) and Sigmoid (purple) activations.

As depicted in Figure 2, a 24-dimension input is received (8 frames for each input type at once) for each timestep. Furthermore, the recurrent representation is concatenated to the encoded data for each timestep. The idea behind this merge lies on the fact that it would not be possible to minimize the value of the multi-output reconstruction loss function with only the low-dimensional representation obtained from the GRU layer. Besides that, increasing the size of this recurrent layer would greatly increase the number of parameters and the time to train the models, requiring more training data as a consequence. It is important to notice that the models consider as input a 4-timestep tensor and after that, the state of the GRU layer is reseted.

The proposed approach employs an autoencoder model for each class of the dataset. The main argument that supports this choice is related to scalability in the sense that the recognition of a novel gesture only requires the training of one new autoencoder model for the novel gesture, without the need for retraining for all instances of the training set, as commonly performed. Other reasons for choosing one autoencoder for each class are listed as follows.

(i) Lower complexity: Since intra-class variation tends to be lower than inter-class, it is possible to generate autoencoder models with lesser parameters, resulting in models with faster training, easier convergence, less prone to overfitting and with a lower training data requirement.

(ii) Specificity: Classes with a higher complexity requirement can be associated to higher capacity models without requiring any retraining of previous models. In addition, since MORA employs a multi-output loss function, it is possible to assign specific loss weights to each class, considering the most relevant aspects for each of them.

(iii) Open-set applicability: Since models tend to minimize the reconstruction error of trained instances, it is expected to obtain a high error for an instance of an unknown class. Therefore, it is possible to associate high error values to non-trained classes, allowing the employment of MORA on open-set applications. Furthermore, MORA presents high similarity with incremental learning methods and could be easily adapted to consider new data without requiring any change on previously trained models.

(iv) Robustness to unbalanced classes: Each model of MORA aims to minimize the reconstruction error for a specific gesture class. Therefore, the only impact MORA suffers from unbalanced datasets (classes with large variance in number of instances) is the fact that the larger classes tend to require models with higher capacity.

4. Experimental Results

To evaluate the proposed approach, experiments are conducted on SKIG [9] and on ChaLearn IsoGD [20] datasets. For both, the protocol described on [22] is applied, making our results comparable to the ones obtained on their research. It is important to highlight that Farneback optical flow [7] was computed over RGB videos from both datasets. In addition, models with the same configuration (and capacity) were used to represent all gestures.

Experimental Setup. The parameters for each layer of the models are determined by tests on the validation set of SKIG dataset. Since the model intends to reconstruct the input data, it is important not to stack layers with a large difference in number of parameters [18] (number of feature maps, for instance). Besides, producing very low-dimensional representations could lead to high reconstruction errors, hindering the minimization of the loss function.

The employment of GRU is justified by the fact that this type of recurrent layer is able to extract relevant information from data that presents a well defined temporal structure. LSTMs are also capable of such, however, they have a larger number of parameters to be adjusted and tend not to behave well when a small set of input data is available [8].

To train the model, the learning rate was set experimentally to 0.001 and all convolutional layers employ ReLU activation, except for the last ones (denoted in purple on Figure 2), which employ the sigmoid activation. The GRU layer uses a sigmoid activation for the output and a hard sigmoid for the recurrence. All models contain 2.7Mi parameters and were trained using a Geforce GTX 1060M.

Datasets. The *Sheffield Kinect Gestures* (SKIG) dataset [9] is composed by $1,080$ RGB and $1,080$ depth videos from 10 different gestures. All gestures are performed by six different subjects, with three hand postures, three different backgrounds and two illumination conditions, resulting in 108 videos for each class of gestures. Tests conducted to evaluate MORA include the split on a 3-fold cross-validation, as specified in [22].

The *ChaLearn Isolated Gestures* (ChaLearn IsoGD) dataset [20] is composed by $47,933$ RGB and depth videos from 249 different gestures. Gestures are performed by 21 different subjects and, differently from SKIG, classes are not balanced. Since ChaLearn IsoGD is composed by some very short videos (with approximately 12 frames), the subsampling described in the previous section produced 12-frame inputs and the video clips are composed by only three frames instead of eight. Furthermore, the architecture showed in Figure 2 needed to have its input adjusted. Instead of receiving videos composed by 24 frames (for each timestep), as shown on Figure 2, the architecture received 9-frame inputs (3 for each modality). In addition, to reconstruct outputs with the same dimensionality of inputs, the first two pooling and last two upsampling layers were adjusted, with their kernels presenting a temporal (un)pooling factor of 3 (e.g., $3 \times 2 \times 2$ instead of $2 \times 2 \times 2$).

The protocol used on ChaLearn experiments considers outcomes on the validation and test subsets, as exposed in [22]. To verify MORA's scalability and its behavior in comparison to a standard literature classification model (C3D [2]), an additional test was conducted considering 50 ChaLearn IsoGD classes (randomly selected).

The C3D network is commonly used as baseline for video information extraction since it contains 3D convolutional layers, which extract spatiotemporal information from videos and presents good results in applications regarding activity and gesture recognition [12, 3]. Furthermore, with the adjustment of the input to act over ChaLearn IsoGD videos, C3D presents a complexity, in terms of number of parameters, which is comparable to the sum of employed MORA models, leading to a fair comparison.

Evaluation on the SKIG Dataset. To evaluate MORA, a first experiment is conducted on the SKIG dataset. In this experiment, ten different models were assembled and trained, representing each class of the dataset. Based solely on the reconstruction error, it was possible to label dataset videos reaching an average accuracy of 93.61%.

In a second experiment, activations of the convolutional layer 10 (shown in Figure 2) were associated to a cheap-to-train classifier. The application of this classifier tends to enforce discriminative characteristics of learned representations. Since the autoencoder models are not discriminative, it is expected that outcomes from a classification model surpass the ones obtained purely with reconstruction error.

It is worthy mentioning that this experiment claims for the application of a very simple and cheap classifier with almost no impact on scalability. Differently from a deep classification model that aims at learning discriminative features over input data, this simple classifier only intends to separate data considering features already learned by autoencoder models, a much faster process. Thus, to correctly produce feature vectors representing videos without any bias, the activations of all ten models were concatenated to compose the vector for each video. Since the chosen layer employs a ReLU activation and models are trained to generate a response only for the class they are trained, the obtained vector is extremely descriptive and sparse. With that, the use of a simple 15-neurons multilayer perceptron classifier was enough to achieve an accuracy of 97.20%.

Table 1 shows the results achieved by several methods on the SKIG dataset. Although simple, MORA reaches high accuracies, being comparable to sophisticated state-of-the-art methods. Since SKIG is a 10-class balanced dataset, the advantages of MORA are not very evident. However, when the complexity of models is considered besides the consequences of training a huge number of parameters, MORA shows to be valuable even when considering the outcomes obtained with only the reconstruction of inputs. In addition, MORA is scalable in terms of number of classes, an advantage in this context. The approaches that surpassed MORA in terms of accuracy present a much higher complexity/capacity besides sophisticated fusion techniques to obtain these outcomes. Furthermore, since the protocol for SKIG dataset does not demand for the computation of dispersion metrics and/or significance tests, it is difficult to assert the superiority of state-of-art methods over the proposed approach.

Evaluation on the ChaLearn Dataset. The experiments on ChaLearn IsoGD dataset intend to evaluate MORA's behavior on a large class-unbalanced scenario. With that, it is possible to analyze whether the reconstruction error of recurrent non-discriminative models provides enough information to separate a large amount of classes, besides the robustness of MORA to class imbalance.

Tables 2 and 3 show MORA outcomes and reference methods on the dataset. For both subsets of the ChaLearn IsoGD (validation and test), MORA outcomes were based on reconstruction error and activations associated to a cheap 50-neurons Multilayer Perceptron classifier. According to

Table 1: Accuracies of different approaches applied to the SKIG dataset.

	Approach	Acc (%)
Results	RGGP+RGB-D [9]	88.70
	4DCOV [5]	93.80
	Depth Context [10]	95.37
	Tung et al. [19]	96.70
	MRNN [13]	97.80
	3DCNN+RNN+CTC [12]	98.60
	Zhang et al. [22]	99.53
Our Results	MORA reconstruction	93.61
	MORA activations + 15N MLP	**97.20**

Table 2: Accuracies on the validation subset of the ChaLearn IsoGD dataset.

	Approach	Acc (%)
Results	Pyramidal C3D. [24]	45.02
	Zhang et al. [22]	58.65
Our Results	MORA reconstruction	50.23
	MORA activations + 50N MLP	**57.76**

Table 3: Accuracies on the test subset of the ChaLearn IsoGD dataset.

	Approach	Acc (%)
Results	Pyramidal C3D. [24]	50.93
	Zhang et al. [22]	62.14
	2SCVN-3DDSN. [6]	67.26
Our Results	MORA reconstruction	56.37
	MORA activations + 50N MLP	**66.16**

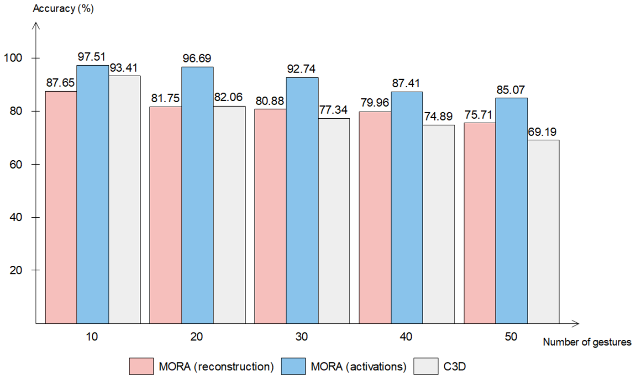

Figure 3: Accuracy on 50 classes of Chalearn.

the results on the validation subsets, MORA presents a very similar accuracy outcome to the method proposed by Zhang [22], which employs a much more complex model with sophisticated fusion of modalities. On the test subset, MORA outperforms Zhang's method, being only surpassed by 2SCVN-3DDSN [6], which employs ensemble learning that integrates Two Stream Consensus Voting Network (2SCVN) and 3D Depth-Saliency Network (3DDSN).

Table 4: Comparison between MORA and a discriminative model (C3D) for tests on ChaLearn.

	MORA	**C3D**
Number of Parameters	2.7M per model	111M[2]
Average Time per Epoch (minutes)	4.89[3]	7.83
Average Time to insert new class (minutes)	75	1263[4]
Time to Train Custom Classifier (minutes)	8 (10 gestures) 13 (20 gestures) 15 (30 gestures) 20 (40 gestures) 24 (50 gestures)	Not required

Scalability Evaluation on the ChaLearn IsoGD. Our last experiment consists on the application of MORA on 50 classes (randomly selected) from the ChaLearn IsoGD dataset. For this, the validation subset of ChaLearn was considered and an increasing number of gestures was employed, with their accuracies being shown on Figure 3. In addition, the accuracy of a widely employed gesture/activity recognition classification model (C3D [2]) is also shown.

According to the results, the accuracy drop obtained with MORA outcomes shows a soft behavior. In contrast, the C3D model shows a high accuracy for the first test (10 gestures), but much lower values on the subsequent ones. This behavior can be associated to: (i) C3D model is a classification (discriminative) network and it is highly impacted by class imbalance, which is more evident when a higher number of classes is considered; (ii) different from MORA, C3D is a fixed-capacity model. Thus, the same architecture was employed for all tests, with no variation on the number of parameters. It is important to notice that MORA with activations employs a custom Multilayer Perceptron classifier, which is susceptible to class imbalance. However, since the features used by this classifier were learned in an unsupervised way, the impact of class imbalance is reduced.

Table 4 lists a set of parameters from MORA and C3D models. According to the table, MORA employs a variable number of parameters depending on the task and presents a low time requirement regarding the addition of a new class (i.e., new gesture). In addition, even presenting higher accuracy outcomes than C3D, MORA is, for most cases, a much less complex model and presents a lower time per iteration for training. Even though MORA's complexity is similar to C3D for the classification of 50 gestures, the advantage of the present approach relies on incremental growth on number of parameters, what improves the capacity of MORA depending on the number of classes. Finally, the cost of inserting a discriminative classifier is also presented, evidencing the low impact on MORA's scalability, since the cost to train this network is very low.

[2] This value is adjusted to receive ChaLearn videos as inputs.

[3] This average value represents the time to train 50 MORA models.

[4] Even with a similar time per epoch, C3D requires more iterations since it contains much more parameters to be trained.

5. Concluding Remarks

This paper presented a gesture recognition approach, the *Multi-Output Recurrent Autoencoders* (MORA), based on the application of multiple recurrent autoencoder models. Tests were conducted on SKIG and on ChaLearn IsoGD datasets, achieving recognition accuracies similar to state-of-art methods. Since MORA does not represent a classification approach, it was expected to obtain lower recognition rates than methods such as the ones proposed by Zhang et al. [22] and Molchanov et al. [12]. However, MORA shows to be valuable for obtaining similar recognition rates utilizing lower capacity/complexity models. In addition, MORA's scalability and robustness to class-imbalance, demonstrated through experiments on ChaLearn IsoGD, represents a positive characteristic of the approach, along the possibility to specify models and loss weights for some classes of a dataset. Finally, since every model is trained in an independent way and each of them is related to a dataset class, MORA presents a straight forward applicability in the context of open-set/incremental learning.

Acknowledgments

The authors would like to thank the National Council for Scientific and Technological Development – CNPq (Grant 311053/2016-5), the Minas Gerais Research Foundation – FAPEMIG (Grants APQ-00567-14 and PPM-00540-17) and the Coordination for the Improvement of Higher Education Personnel – CAPES (DeepEyes Project).

References

[1] I. L. O. Bastos, M. F. Angelo, and A. Loula. Recognition of static gestures applied to brazilian sign language (libras). In *SIBGRAPI*, pages 305–312, 2015.

[2] D. T. L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE ICCV*, pages 4489–4497, 2015.

[3] C. Cao, Y. Zhang, Y. Wu, H. Lu, and J. Cheng. Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules. In *IEEE ICCV*, 2017.

[4] J. Chung, C. Gülçehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, 2014.

[5] P. Cirujeda and X. Binefa. 4dcov: A nested covariance descriptor of spatio-temporal features for gesture recognition in depth sequences. In *3DV*, pages 657–664, 2014.

[6] J. Duan, S. Zhou, J. Wan, X. Guo, and S. Li. Multi-modality fusion based on consensus-voting and 3d convolution for isolated gesture recognition. *CoRR*, 11 2016.

[7] G. Farnebäck. Very high accuracy velocity estimation using orientation tensors parametric motion and simultaneous segmentation of the motion field. In *IEEE ICCV*, pages 171–177, 2001.

[8] K. Greff, R. Srivastava, J. Koutník, B. Steunebrink, and J. Schmidhuber. LSTM: A search space odyssey. *CoRR*, 2015.

[9] L. Liu and L. Shao. Learning discriminative representations from rgb-d video data. In *IJCAI*, pages 1493–1500, 2013.

[10] M. Liu and H. Liu. Depth context. *Neurocomput.*, 175:747–758, Jan. 2016.

[11] S. Mitra and T. Acharya. Gesture recognition: A survey. *IEEE Transactions on Systems, Man and Cybernetics*, 37(3):311–324, 2007.

[12] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks. In *IEEE CVPR*, pages 4207–4215, 2016.

[13] N. Nishida and H. Nakayama. Multimodal gesture recognition using multi-stream recurrent neural network. In *7th Pacific-Rim Symposium on Image and Video Technology - Volume 9431*, pages 682–694, 2016.

[14] V. Pavlovic, R. Sharma, and T. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):677–695, 1997.

[15] L. Pigou, A. Oord, S. Dieleman, M. Herreweghe, and J. Dambre. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *CoRR*, 2015.

[16] A. Shahroudy, J. Liu, T. Ng, and G. Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *IEEE CVPR*, 2016.

[17] C. Souza, F. Pádua, V. Lima, A. Lacerda, and C. Carneiro. A computational approach to support the creation of terminological neologisms in sign languages. *Computer Applications in Engineering Education*, 2018.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. *ArXiv e-prints*, Dec. 2015.

[19] P. Tung and L. Ngoc. Elliptical density shape model for hand gesture recognition. In *5th Symposium on Information and Communication Technology*, pages 186–191, 2014.

[20] J. Wan, S. Li, Y. Zhao, S. Zhou, I. Guyon, and S. Escalera. Chalearn looking at people RGB-D isolated and continuous datasets for gesture recognition. In *IEEE CVPR*, 2016.

[21] S. Xu and Y. Xue. A long term memory recognition framework on multi-complexity motion gestures. In *ICDAR*, volume 01, pages 201–205, 2017.

[22] L. Zhang, G. Zhu, P. Shen, J. Song, S. A. Shah, and M. Bennamoun. Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition. In *IEEE ICCV*, 2017.

[23] H. Zhou and Q. Ruan. A real-time gesture recognition algorithm on video surveillance. In *8th International Conference on Signal Processing*, volume 3, 02 2006.

[24] G. Zhu, L. Zhang, L. Mei, J. Shao, J. Song, and P. Shen. Large-scale isolated gesture recognition using pyramidal 3d convolutional networks. In *ICPR*, pages 19–24, 12 2016.