Journal of Engineering and Applied Sciences 13 (21): 8949-8954, 2018
ISSN: 1816-949X

Convolutional Neural Training for Robotic Control Through Hand Gestures

[1]M. Robinson Jimenez, [2]S. Paola Nino, [1]Oscar F. Aviles and [3]Diana Ovalle
[1,3]Mechatronics Engineering Program, Faculty of Engineering, Militar Nueva Granada University, Bogota, Colombia
[2]National Polytechnic Institute, Mexico D.F, Mexico
[3]Distrital Fco. Jose de Caldas University, Faculty of Engineering, Bogota, Colombia

Abstract: This study presents the training of a convolutional neural network to identify different control signals made by hand, that allow to command a robotic mobile. Initially a database of 4000 images is established regarding the different control signals for the manipulation of the mobile, corresponding to 10 different users and after this the base structure of the convolutional neural network and the results of its training are determined. The robotic control algorithm was validated by means of navigation tests performed by 5 different users to those employed in the training stage where a percentage of accuracy was obtained to perform linear paths on average of 93.2% and for non-linear paths of 79%. Training algorithms for convolutional neural networks have not been evaluated in robotic navigation control tasks for transporting objects.

Key words: Deep learning, convolutional network, robotic control, hand gestures, human computer interaction, convolutional, percentage

INTRODUCTION

Robotics, since, its conception as tools to develop heavy work has remained in constant evolution. The way to implement and instrument robotic agents from fixed to mobile goes hand in hand with the technological development of sensor systems such as cameras and the development of pattern recognition algorithms such as neural networks.

The integration of the information of an image, obtained by means of a camera and the discrimination of the objects within that image, find in modern techniques such as convolutional neural networks a means to achieve the interaction of a robotic agent with its surroundings.

The Convolutional Neural Networks (CNN) are part of the so-called Deep Learning (DL) algorithms which seek to emulate the human brain's learning form based on the multilayer concept. The state of the art in DL-based developments cover applications in a variety of fields such as time series data modeling (Langkvist *et al*., 2014) and prediction of pathologies caused by solar radiation (Barrera *et al*., 2015). In the character recognition part, the free handwriting recognition applications (Walid and Lasfer, 2014) as well as other similar patterns such as those presented by Qi *et al.* (2014), Zhou *et al.* (2013) and Wang *et al.* (2014) are highlighted including speech recognition. By Angelova *et al.* (2015) DL techniques are used to carry out the detection of pedestrians in an image, with a view to the development of autonomous vehicles, which can be complemented with the work presented by Chen *et al.* (2014), for vehicle detection from satellite images. With respect to this type of applications that propose a form of human-machine interaction by Luo *et al.* (2017) the training of an 8-layer deep convolutional neural network for vehicular recognition is also presented, also applicable to interaction with pedestrians.

It can be observed specifically that the applications that focus on the detection and recognition of patterns in images under DL are developed through CNN's. Other examples of specific state of the art in CNN are presented by Zhang and Zhang (2014) where a face detection system is presented for the identification of faces, taking a base of 117 thousand faces subject to changes of pose, expression and illumination in order to strengthen the training of the convolutional network for which 15 subcategories are used in which a possible desired face can be found. By this same way are applications of face detection for purposes such as identifying states of numbness in a driver (Dwivedi *et al*., 2014), recognition of expressions (Song *et al*., 2014) and recognitions of features of facial beauty (Gan *et al*., 2014).

Some previous work sets the training of CNNs for the detection of hand gestures (Tompson *et al*., 2014;

Corresponding Author: M. Robinson Jimenez, Mechatronics Engineering Program, Faculty of Engineering, Militar Nueva Granada University, Bogota, Colombia

Jacobs *et al*., 2016; Strezoski *et al*., 2016) which have also been previously tested with image processing algorithms (Malviya and Chawla, 2017). However, there are no applications that integrate such recognition through CNN to control robots or human-machine interaction applications. In this area of development there are few applications of CNNs where the closest one is oriented in the estimation of the possible location of a robotic gripper for object grasp (Wang *et al*., 2016).

In this resaerch the training of a convolutional network for the identification of robotic commands derived from hands signals is presented, presenting the characteristics and results of this training where part of the founding for this purpose is exposed by Zeiler and Fergus (2014). The results of signal recognition for the navigation of a robotic mobile in two types of path, one linear and the other non-linear are also presented.

MATERIALS AND METHODS

CNN training: A convolutional neural network seeks to set a group of convolution filters which are trained by a training image database and the general structure of the network. Said structure is composed of a series of consecutive layers of convolution relu-pooling or variations there of in a characteristic extraction stage and followed by another classification. Filters make up a group of feature maps where each map is obtained by repeated application of the filter function across sub-regions of the entire input image, according to Eq. 1:

$$h_j^n = \max\left(0, \sum_{k=1}^{k} h_k^{n-1} * w_{kj}^n\right) \quad (1)$$

Where:

h_j = Determines the output characteristics map
h_k = The input which for the initial stage will correspond to the characteristics of the image
k = 1 = If it is grayscale or 3 if it is in color, finally wk corresponds to the kernel of convolution

For the case of the identification of control signals in this resarch, we opt for the training of a convolutional neural network whose graphic structure is represented in Fig. 1. The conventional techniques of image processing require different stages to achieve the segmentation of the hand in an image, once obtained this segmentation it is necessary to use some technique of pattern recognition. Training through the convolutional neural network covers each of these aspects.

The neuronal structure presented has a convolution input for color images, i.e., three-dimensional with a known Height (H) and Width (W) and equal to 64 pixels. The training hyperparameters to set the architecture of the convolutional neural network are: Stride (S) or displacement of the filter to the input volume. Padding (P) or lateral filling with zeros, the size of the Filters set (F), which according to the stride ends in W.

In this way, the result of the convolution enters a layer called RELU (rectified linear unit) which is an activation function layer. The output volume dimensions of the n-layer of the convolution to the RELU are calculated by Eq. 2:

$$W_2 = \frac{(W_1\text{-}F+2P)+1}{S}$$
$$H_2 = \frac{(H_1\text{-}F+2P)+1}{S} \quad (2)$$

The pooling layer operates independently from the convolution layer and progressively reduces the size of the layers by means of the maximum or average methods, for this case the maximum method is used which is established by Eq. 3 where h_j determines the new map of characteristics of this stage, based on the previous one and its own value of stride:

$$h_j^n(x,y) = \max \bar{x} \in N(x), \bar{y} \in N(y) h_j^{n-1}(\bar{x},\bar{y}) \quad (3)$$

As an object of recognition, a series of hand signals are established which can be seen in Fig. 2. The ten

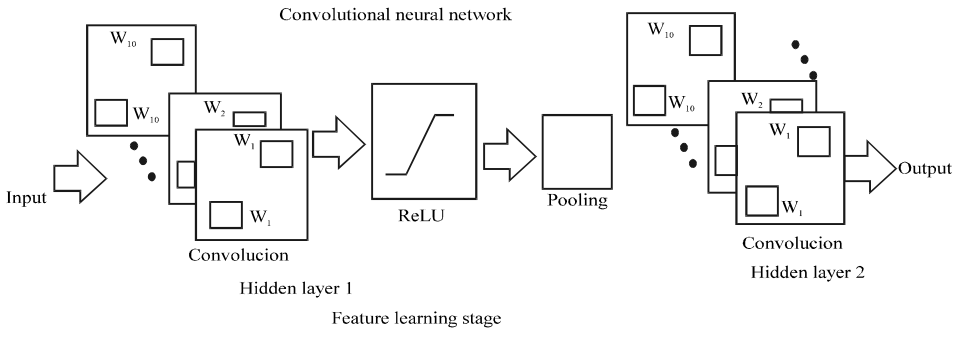

Fig. 1: Layered structure of the CNN used

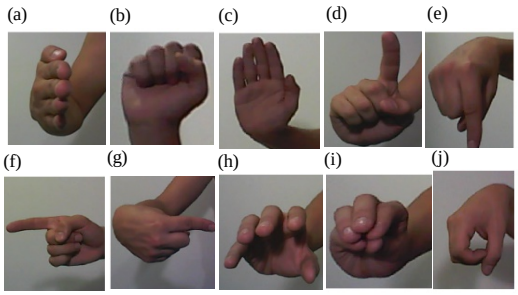

Fig. 2: a-j) Hand gestures to discriminate throug CNN

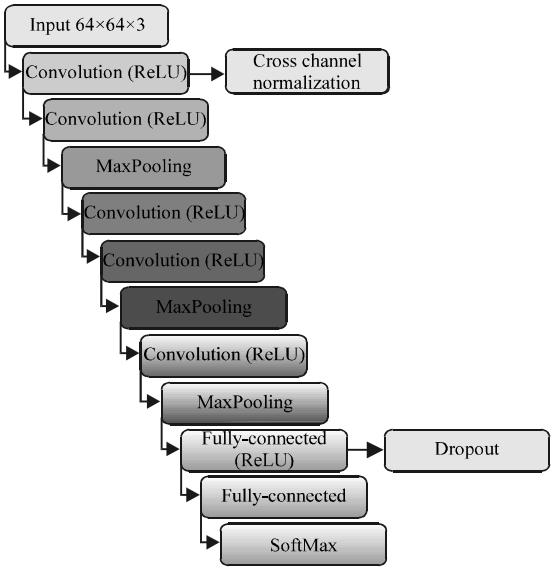

Fig. 3: CNN architecture used

gestures associated with the movements to be used in the robotic agent are: forward, backward, stop, up and down, from left to right at the bottom of the figure. For each case, which are observed in that order from left to right in the upper part of the figure and the gestures of the right and left movements, open, close and rotate gripper in order the database consisted of 400 photos per movement with 10 different users for a total of 4000 images with a size of 640×480 pixels of which half was used for training and the other half for validation.

The network training is established in function of the uniformity in the size of each image of the database used for the training and for validation, so they are rescaled to a scale of 64×64 (H×W) pixels from the original image in order to balance the number of samples and the size of each which directly affects the architecture of the convolutional network and processing times. For this image size, 20 convolution filters are used in the initial or

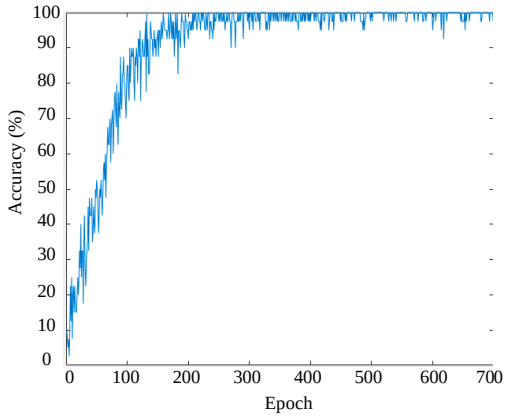

Fig. 4: Training response for calculation of epochs

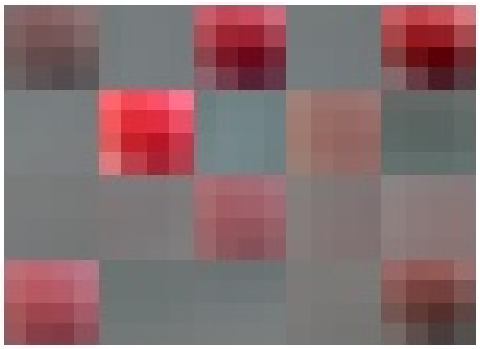

Fig. 5: Learned convolution filter of the first layer

input stage with a size of 4×4 (F = 4) each in conjunction with the previously mentioned hyperparameters set in: P = 2, S = 1.

Figure 3 illustrates the architecture of the convolutional neural network used which is appreciated is of deep type.

Training results: In Fig. 4, it is possible to observe the performance of the network where in the range of 250-700 epochs a similar behavior is <95%, choosing 643 as the number of training epochs of the network hence to the 100% accuracy stability in the response. In Fig. 5, one of the learned convolution filters of the first layer is shown.

The result of the prediction of each gesture indicated by the hand which in turn is representative of a control command of a robotic mobile is tabulated in the confusion matrix shown in Fig. 6. The final average result shows a 71% effectiveness of the network where the signals with the highest degree of accuracy are the up (4) and the right (6) with 95% denoted in bold while the lower ones are the back (2) and close gripper (9) with 40 and 55%, respectively.

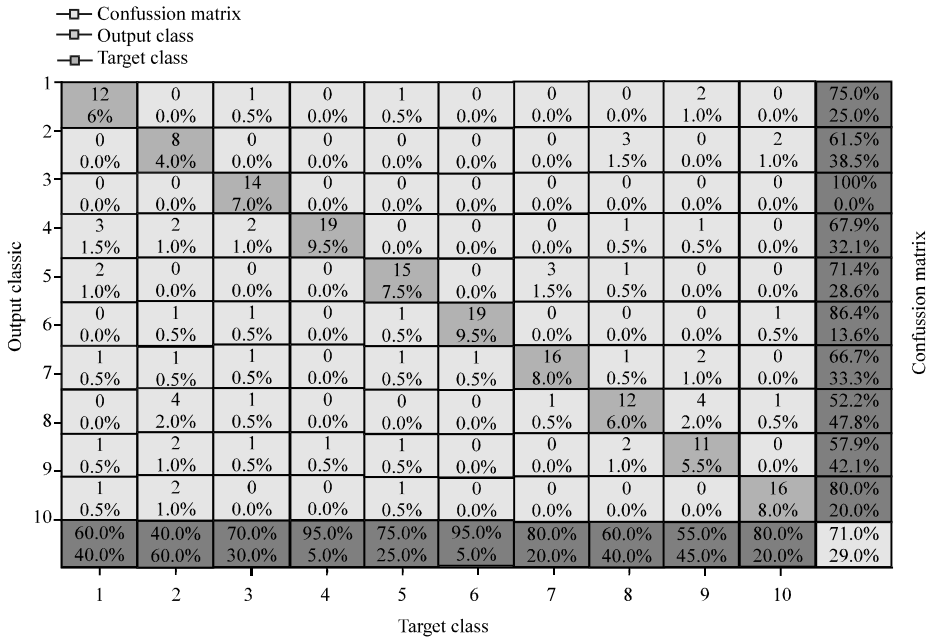

Fig. 6: Final confusion matrix

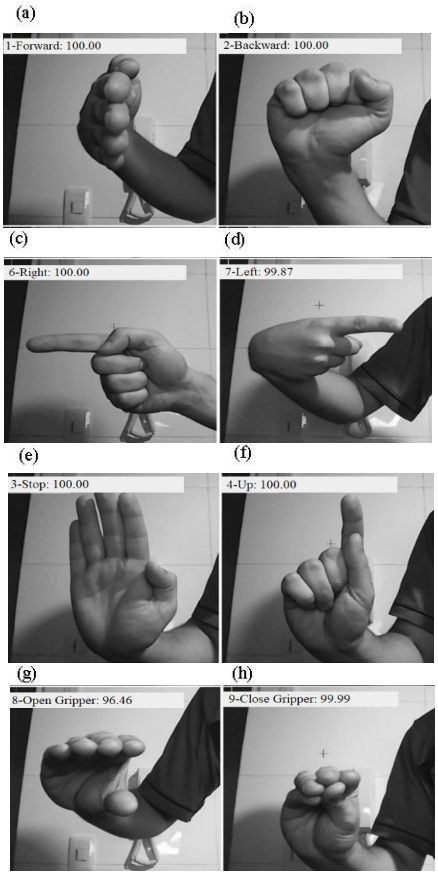

Fig. 7: Network response in real environment

The result of the prediction in a real test environment of each gesture is illustrated in Fig.7. It is possible to appreciate the clear recognition of each of the signals, so that, the recognition rate is higher than 96% in the worst case which corresponds to the down gesture (5), i.e., a particular case is presented under a controlled environment in which the trained network operates satisfactorily, despite the fact that the average result can be considered low.

RESULTS AND DISCUSSION

To validate the functionality of the control signals set by the hand, a robotic mobile of 4 degrees of freedom is used which is seen in Fig. 8, so that, it is able to reproduce the movements associated to the database of the training images. The training of the convolutional neural network is performed from MATLAB® where a webcam capture interface is used to get the user's hand, then the image is submitted to the trained network for recognition of the action signal.

Once the command signal type is acknowledged, the associated decimal number (from 1-10) is sent to a serial port of the computer to which it is connected an Arduino Uno microcontroller card that transmits said value by means of radiofrequency using an XBee pro series 2 device.

The robotic mobile acquires the signal by this same route and according to the received value the DC motors

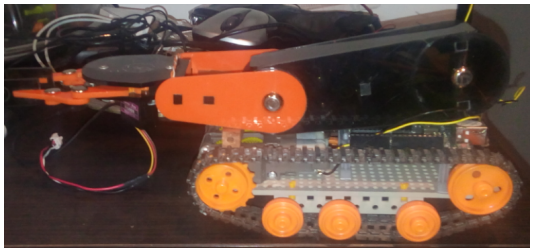

Fig. 8: Robotic mobile

Table 1: Robotic control results

Input signal	Number trials	Success	Fail
Forwards	25	22	3
Stop	25	19	6
Up	25	21	4
Left	25	23	2
Right	25	24	1
Open gripper	25	17	8
Close gripper	25	15	10

Table 2: Results confusion matrix

Classification	Open gripper	Close gripper
True positive	91	84
False positive	8	16

Table 3: Results of robotic mobile paths

Paths/user	1	2	3	4	5
1					
Time (sec)	56	49	59	55	52
Accuracy (%)	95	93	94	90	94
2					
Time (sec)	97	91	85	95	112
Accuracy (%)	78	77	81	76	83

of displacement and rotation of the mobile operate or generates the respective signals to manipulate the servomotors of location and action of the gripper.

Table 1 illustrates the results obtained in the evaluation of the final network trained on the action of the mobile. A base of 25 snaps was taken for each command signal to the robot where the number of times each signal was executed correctly and those not.

It can be seen that the highest values at the exit of the network correspond to the actions of forward, stop, right turn and left turn. When analyzing the hand gestures associated with these actions, it can be seen that they have the most discriminating characteristics, for example, for the case of opening and closing clamp the frontal similarity of the action generates confusion in the detection by the network. Observed this fact, a temporal identification relation is set, i.e., the decision of opening and/or closure of clamp is made in function of the detection of this signal during three consecutive catches, and the results obtained are tabulated in Table 2 where a notable improvement in detection is evidenced.

The functionality of the algorithm and the hardware developed was evaluated by mobile displacement tests in object grabbing and storage tasks. Two paths were proposed which are illustrated in Fig. 9 with a Start Point

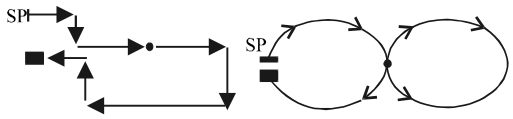

Fig. 9: Paths of test

(SP) and end point denoted by the black box. These were implemented with guide lines on the ground as a reference to a group of 5 users. The time it took for each user to complete the path was evaluated, taking an object along the way as well as evaluating the accuracy when leaving the object at a defined point which was calculated by the difference between the distance from the geometric center of the object at the point of discharge to the target point. The results obtained are presented in Table 3.

It can be observed in Table 3, that with non-linear paths such as path 2, the complexity of the mobile control increases, requiring more time and reducing the accuracy reached. It can also be seen that, in the developed algorithm, there is no significant incidence in the control of the mobile when changing users, since, in each case the test is successfully performed, although with different degrees of precision which is observed as it relates to the time spend by the user and his ability.

CONCLUSION

A neural network of the convolutional type was designed and trained for discrimination of a series of command signals for a robotic agent. Having as training inputs a considerable database (2000), the network efficiency was 71% which seems low. However, the validation tests and the temporary retention adjustment for the less favorable cases of recognition, allowed a high level of control of the mobile by different users.

The selection of the control signals plays a fundamental role in their discrimination, it is observed that at least one clearly distinguishable characteristic is required to ensure their identification which is the problem presented in the signals backward (2) and close gripper (9) because they are similar will present confusion of classes.

The validations made allow extending the field of application of the algorithm developed to applications of robotic control of displacement such as robotic wheelchairs for disabled people and even with some degree of paraplegia.

ACKNOWLEDGEMENT

The research for this study was supported by Distrital Fco. Jose de Caldas University and Davinci Research Group of Nueva Granada Military University.

REFERENCES

Angelova, A., A. Krizhevsky and V. Vanhoucke, 2015. Pedestrian detection with a large-field-of-view deep network. Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA), May 26-30, 2015, IEEE, Seattle, Washington, USA., ISBN:978-1-4799-6924-1, pp: 704-711.

Barrera, P.J.F., A.D. Hurtado and J.R. Moreno, 2015. Prediction system of erythemas for phototypes i and ii, using deep-learning. Vitae, 22: 189-196.

Chen, X., S. Xiang, C.L. Liu and C.H. Pan, 2014. Vehicle detection in satellite images by hybrid deep convolutional neural networks. IEEE. Geosci. Remote Sens. Lett., 11: 1797-1801.

Dwivedi, K., K. Biswaranjan and A. Sethi, 2014. Drowsy driver detection using representation learning. Proceedings of the 2014 IEEE International Advance Computing Conference (IACC), February 21-22, 2014, IEEE, Gurgaon, India, ISBN:978-1-4799-2573-5, pp: 995-999.

Gan, J., L. Li, Y. Zhai and Y. Liu, 2014. Deep self-taught learning for facial beauty prediction. Neurocomputing, 144: 295-303.

Jacobs, K., M. Ghasiazgar, I. Venter and R. Dodds, 2016. Hand gesture recognition of hand shapes in varied orientations using deep learning. Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists, September 26-28, 2016, ACM, Johannesburg, South Africa, ISBN:978-1-4503-4805-8, pp: 17-17.

Langkvist, M., L. Karlsson and A. Loutfi, 2014. A review of unsupervised feature learning and deep learning for time-series modeling. Pattern Recognit. Lett., 42: 11-24.

Luo, X., R. Shen, J. Hu, J. Deng and L. Hu *et al*., 2017. A deep convolution neural network model for vehicle recognition and face recognition. Procedia Comput. Sci., 107: 715-720.

Malviya, A.K. and M. Chawla, 2017. Effects of different color models in hand gesture recognition. Indian J. Sci. Technol., 8: 1-8.

Qi, Y., Y. Wang, X. Zheng and Z. Wu, 2014. Robust feature learning by stacked autoencoder with maximum correntropy criterion. Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 4-9, 2014, IEEE, Florence, Italy, ISBN:978-1-4799-2893-4, pp: 6716-6720.

Song, I., H.J. Kim and P.B. Jeon, 2014. Deep learning for real-time robust facial expression recognition on a smartphone. Proceedings of the 2014 IEEE International Conference on Consumer Electronics (ICCE), January 10-13, 2014, IEEE, Las Vegas, Nevada, USA., ISBN: 978-1-4799-1289-6, pp: 564-567.

Strezoski, G., D. Stojanovski, I. Dimitrovski and G. Madjarov, 2016. Hand gesture recognition using deep convolutional neural networks. MSc Thesis, Saints Cyril and Methodius University of Skopje, Skopje, Republic of Macedonia.

Tompson, J., M. Stein, Y. Lecun and K. Perlin, 2014. Real-time continuous pose recovery of human hands using convolutional networks. ACM. Trans. Graphics, 33: 1-10.

Walid, R. and A. Lasfar, 2014. Handwritten digit recognition using sparse deep architectures. Proceedings of the 2014 9th International Conference on Intelligent Systems: Theories and Applications (SITA-14), May 7-8, 2014, IEEE, Rabat, Morocco, ISBN: 978-1-4799-3568-0, pp: 1-6.

Wang, Y., A. Narayanan and D. Wang, 2014. On training targets for supervised speech separation. IEEE. ACM. Trans. Audio Speech Lang. Process., 22: 1849-1858.

Wang, Z., Z. Li, B. Wang and H. Liu, 2016. Robot grasp detection using multimodal deep convolutional neural networks. Adv. Mech. Eng., Vol. 8,

Zeiler, M.D. and R. Fergus, 2014. Visualizing and understanding convolutional networks. Proceedings of the European Conference on Computer Vision, September 6-12, 2014, Springer, Zurich, Switzerland, pp: 818-833.

Zhang, C. and Z. Zhang, 2014. Improving multiview face detection with multi-task deep convolutional neural networks. Proceedings of the 2014 IEEE Winter Conference on Applications of Computer Vision (WACV), March 24-26, 2014, IEEE, Steamboat Springs, Colorado, USA., ISBN:978-1-4799-4984-7, pp: 1036-1041.

Zhou, S., Q. Chen and X. Wang, 2013. Active deep learning method for semi-supervised sentiment classification. Neurocomputing, 120: 536-546.