Human Action Recognition in Realistic Scenes Based on Action Bank

Carlos Martínez, Marcos Baptista, Cristina Losada, Marta Marrón *, and Valeria Boggian.

GEINTRA Group, University of Alcalá.
http://www.geintra-uah.org
{carlos.martinez,marcos.baptista,losada,marta,valeria.boggian}@depeca.uah.es

Abstract. During the last decades topics such as video analysis and image understanding have acquired a big importance due to its inclusion in applications such as security, intelligent spaces, assistive living and focused marketing. In order to validate all related works different datasets have been distributed within the research community: CAVIAR, KTH, Weizmann, INRIA or MuHAVI are some of the most well-known examples, but in most cases these datasets have not been created for the surveillance application in realistic scenes of interest. Within this context, here we present a work that implements a solution for multiple persons' action recognition in monocular video sequences, focused on surveillance applications. Besides, it is also presented a newly created dataset with realistic scenes specifically designed for commercial applications. Development and results of the proposed algorithm and its validation, both within well-known datasets as CAVIAR and KTH and within the one ad-hoc generated for the applications of interest, are discussed in the paper.

Keywords: Action Bank, activity recognition, video-surveillance, monocular RGB image processing

1 Introduction

The current development of audio and video processing technologies applied to cognitive systems allows automating more complex tasks and in a more accurate way. Nowadays several efforts have been done in order to generate cognitive systems whose aim is to analyse the different events that happen to humans in their typical environments. This process is named scene understanding. Scene understanding techniques are thus applied to accomplish human behaviour or activity analysis, normally based on sequences of human actions.

* This work has been supported by the Spanish Ministry of Economy and Competitiveness under project SPACES-UAH (TIN2013-47630-C2-1-R), and by the University of Alcalá under projects DETECTOR (CCG2015/EXP-019) and ARMIS (CCG2015/EXP-054).

Action recognition in video sequences is an area deeply explored due to its initial application to automatic film transcription, that now rebirths in this human behaviour analysis trending application. As an example of such application, we should at least remark the following: biometrics and surveillance [1], prediction of potential risk situations (e.g. capacity excess in a concert hall), social behaviour analysis (e.g. personalized service and marketing [2], and psychological or demographic analysis [3]).

Within this context, in this paper we present a work that implements a solution for human action recognition in RGB video sequences, specifically focused on surveillance applications. Development and results of the proposed algorithm, as well as its validation, both with a well-known dataset as KTH and also with a new one generated ad-hoc for such applications, are discussed in the paper.

2 State of the Art

Being the tackled topic such actual and interesting, there are many relevant publications, just among the last five years scientific literature, that present different approaches and even resume some of the most important ones [4] from different perspectives.

As a first approach to analyse this state of the art, the different solutions can be divided into two groups depending on the technology used. In the last years, there is an increasing literature production of RGBD based solutions [5], [6], [7], [8], [9]. Although this prominent analysis trends to grow during the next decade thanks to the cheaper technology [10], [11], it can be clearly stated that the RGB based solution is the most explored and, at the same time, fruitful one [12], [4], [13], [14], [15], [16]. Besides, this last one appears to be more challenging, as it works with poorer information.

The proposals for human behaviour analysis in images can also be organized according to the very common parameter of using or not a body model, being model based solutions [5], [17], [6], [4], [14], [8], [9] more extended that modeless ones [12], [18], [19].

In this open discussion, the off-the-shelf proposal hereby presented relies on modeless solutions due to the following reasons: algorithms based on a model, either silhouettes [20] or body part detectors (BDPs) [21], need a preprocessing task in which the model is obtained, and are thus more dependent on this preprocessing robustness and accuracy. Besides, modeless solutions have a lower context awareness, as they rely on a priori offline training task that could include the diverse contextual situations related to the specific desired application (as in [22]).

Finally, the state of the art analysis can also be organized according to the behavioural hierarchical level that is obtained. Thus, a first group includes proposals that obtain lower level of behavioral cues, i.e. the simple action being performed, such as running, walking, steadying, sitting, and so on [5], [6], [4], [8], [9], that we call semantic attribute of the scene. A second group includes system proposals that give high level behavioural cues, i.e. complex activities

that normally include two or more of the low level ones, such as playing football or attending in a shop [12], [17], [14], and, in some cases, that include interaction between two or more persons, such as fighting or chatting [23], [13].

As an off-the-self proposal focused on indoor surveillance applications, the solution discussed hereafter would be in the context of low-level behaviour analysis.

Table 1 summarizes the cited works, organized in different groups according to the analysed characteristics.

Sensor		Body model		Behavioural level	
RGB	RGBD	Model based	Modeless	Actions	Activities
[12], [4], [13], [14], [15], [16]	[5], [6], [7], [8], [9]	[5], [17], [6], [4], [14], [8], [9]	[12], [18], [19].	[5], [6], [4], [8], [9]	[12], [17], [14]

Table 1. Classification of the different works according to the analysed characteristics.

In computer vision, it is common to use the available datasets, in order to quantitatively compare the obtained results. There are several datasets that are widely used for human behaviour analysis [24], the most interesting ones, as well as its main characteristics are summarized in table 2.

Dataset	Hieralchical level	Actions	Resolucion	Background	out/indoor
CAVIAR [25]	activities	6	384x288	complex	indoor
INRIA [26]	actions	13	390x291	simple	indoor
KTH [27], [28]	actions	6	160x120	simple	both
MuHAVI [29]	both	17	720x576	complex	indoor
UCF sports [30]	Both	9	720x480	complex	both
Weizmann [31]	actions	10	180x144	simple	outdoor

Table 2. Comparison of different datasets features.

3 Action Recognition

The goal of the developed system is to recognize different everyday human actions, from video sequences recorded by a RGB camera. To achieve this goal, it is necessary to process a set of features from video sequences. These features are used to train a classifier so as to be able to detect various kind of human activities. Once trained the system, another set of sequences will be used for system validation, ensuring in addition a LOPO ("Leave One Person Out") validation scheme. Thus, it can be obtained different quality metrics.

Figure 1 shows a general block diagram including the different parts of the human action recognition system. The input and output on each stage is also shown.

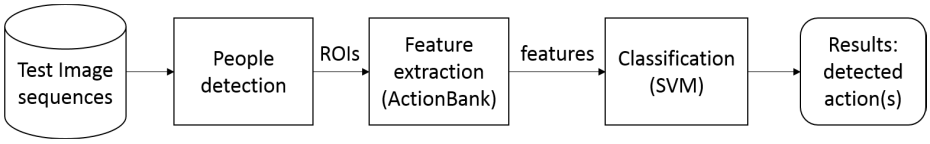

Fig. 1. General block diagram of the action recognition process.

As can be seen in figure 1, the input of the proposed algorithm is an image sequence. This sequence should include one or more people executing the trained actions.

Since in this work we have considered the possibility of detecting more than one action in each image, a people detection stage have been added before the feature extraction. In this stage, we use the well known HOG descriptor in order to detect people in the scene, as well as a people tracker based on a Kalman Filters bank. Once a person has been detected in several consecutive images, the region of interest (ROI) around it in the sequence is sent to the next stage.

Next, a set of descriptors is obtained using Action Bank [12]. We use Action Bank descriptors, instead of other well known ones such as dense point trajectories [32], or dense 3D gradient histograms [33], because it allows obtaining a high-level representation of activities in RGB as well as RGBD video sequences.

Action bank is based on a bank of individual action detectors (previously built using 205 action templates as explained in [12]). The output of detectors are transformed into a feature vector by volumetric max-pooling.

Finally, a Support Vector Machine (SVM) [34], [35] classifier is used for activity recognition, using a one-against-one approach for the multi-class classification process.

The actions that have to be detected strongly depend on the selected image dataset. Generally, the available datasets (see table 2) include simple activities, but they can be not interesting in the context of this work. Since the global aim is action detection for behaviour analysis in surveillance scenarios, from the point of view of the authors there are more important detecting everyday actions such as walking, running or falling down than others such as hand waving or hand clapping. Because of that, the developed system has been trained in order to be able to recognize four different realistic and everyday human actions: walking, running, falling and sitting but new actions can be easily added. Moreover, some results that includes KTH dataset actions are shown in section 5.

4 Dataset for Human Action Recognition

Although there are several dataset for human action recognition available, none of them matches all the requirements of this work. Because of that, we have created a novel dataset that is briefly described in this section.

The created dataset consist on about 280 different indoor image sequences, recorded in the Polytechnics School of the University of Alcalá. These sequences includes four different everyday human actions: walking, running, falling down and sitting down. These actions have been chosen because there are the most interesting in the context of this work, but we are planning to increase the number of actions in the near future.

All the sequences have been recorded using a high definition commercial camera, the GoPro HERO4, witch a 1280x720 pixel resolution and 50 fps. It is situated in an high position in order to reduce the occlusions, and the recorded scene correspond to a section of the first floor of the Polytechnics School of the University of Alcalá. Figure 2 shows the recorded area, as well as the region of interest.

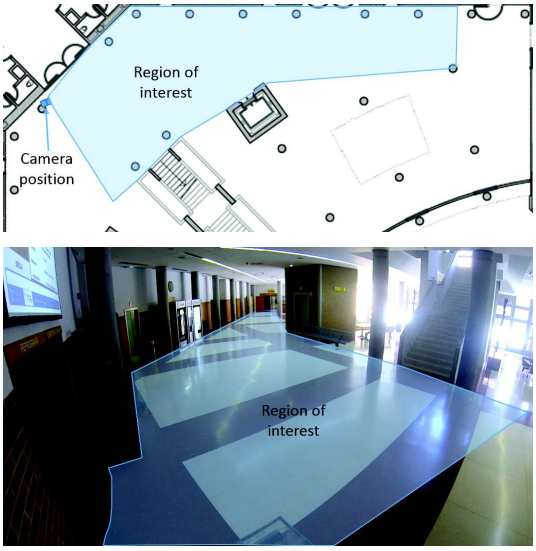

Fig. 2. Recording scenario and region of interest.

As can be seen in figure 2, the background scene is composed for a corridor. On the right side of it there is a stair between two columns, an elevator and some tables. It is worth to highlight that in the chosen scenario have important changes in the illumination depending on the hour of the day that can enrich the dataset with more realistic images (because the illumination is not controlled, and the background can change).

The recorded sequences include different 13 people (called users from this moment) performing one of the defined actions. Table 3 shows the number of sequences for each action as well as the number of different people performing it.

Action	People	Sequences
Walk	13	72
Run	12	75
Sit down	12	68
Fall down	10	58

Table 3. Number of sequences and people performing any of the actions.

There are two kind of sequences: the first one are simple scenes that includes only one person in each image, performing one of the actions, and the second group consists on sequences that include several people, performing different actions simultaneosly. An example of any kind of sequence is shown below. Figure 3shows four consecutive images corresponding to one person falling down. Whereas in figure 4 can be seen an example of images belonging to a complex sequence that includes four different people.

Fig. 3. Example of images corresponding to the falling sequence in the created dataset.

Any of the sequences have been labelled as shown in figure 4 in order to have a ground truth that includes including the position of the person in the image and the performed action.

As explained before, the objective of this contribution is to have a ground truth for testing the developed system and the rest of the scientific community proposals related to video-surveillance applications, including other classes (two new kind of actions: sitting down and falling down) normally present in such applications, and within a realistic scenario.

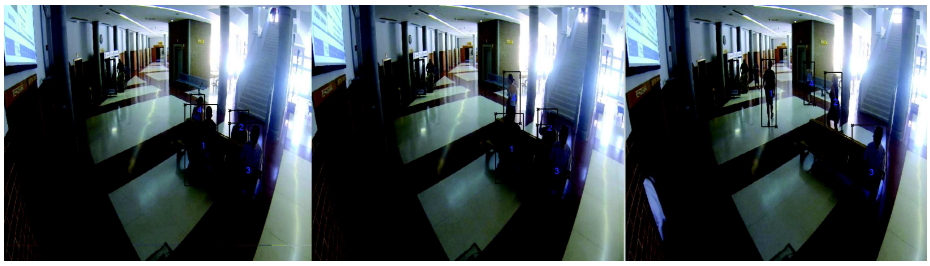

Fig. 4. Example of images and labels corresponding to a complex sequence including different people performing actions.

5 Experimental Results

We have tested and validated the developed system using the well known dataset KTH [27], [28], and an ad-hoc generated one which contains sequences with different persons performing common actions in a realistic scenario.

KTH dataset includes .avi videos recorded with a 25 FPS camera and a resolution of 150x120 pixels. They show six different types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed by 25 subjects in four different scenarios. Some images corresponding to different actions from KTH dataset are shown in figure 5.

Fig. 5. Different examples of KTH images.

5.1 Experimental Tests Using KTH Dataset.

In this section we present the results obtained testing the activity recognition proposal within KTH dataset. The aim of this first test is to compare these results with the ones obtained with the ones presented in the Actionbank reference publication [12].

Different tests, varying the configuration parameters of the LIBSVM classifier [36], were performed, giving the results in table 4. From this table it can be concluded that the recognition task resulting accuracy depends on the type of Support Vector Classification (SVC): the classical soft-margin SVM C-SVC [37], or the nu-SVC [38] that introduces a new parameter $\nu \in (0, 1]$ that is an upper bound on the fraction of training errors and a lower bound of the fraction of

support vectors; as well as on the Kernel (linear, polynomial, radial basis fx, sigmoid) used in the classification process. The values of the different parameters for each Kernel have been obtained empirically.

The obtained results indicate that the best configuration is obtained using a C-SVC SVM classifier with a linear kernel, giving an accuracy of 98.1481%.

SVM Type	Kernel	Accuracy
C-SVC	linear	98.1481 %
	polinomial	95.8330 %
	radial basis fx	20.8330 %
	sigmoid	16.6667 %
nu-SVC	linear	94,9074 %
	polinomial	87,9630 %
	radial basis fx	82,4074 %
	sigmoid	16,6667 %

Table 4. Accuracy obtained for the KTH dataset as a function of the SVM setting.

In figure 6, we finally present the confusion matrix obtained with that best configuration. Most samples are correctly classified. We obtained a 100% accuracy classifying boxing, jogging, running and walking activities, although there is a small confusion between hand waving and hand clapping classes (8.33% and 2.78% respectively).

Fig. 6. Confusion matrix obtained from the KTH dataset.

5.2 Experimental Tests Using the Created Dataset

In this section we present the results obtained with the activity detecting proposal using the database generated within this work. This test allow us to prove the system with a completely different environment and conditions. Furthermore these tests include two more realistic new classes to be detected (falling and sitting) and correctly classified.

Within the set of tests performed using sequences of the database generated for this work (both training and test), we obtained an accuracy of 87.5 %. This is a promising result, especially considering that some of the detected classes are relatively similar (running/walking and sitting/falling) and the small size of the training set (about 80 sequences).

The global confusion matrix obtained is shown in figure 7. As it can be there noticed, it appears a little confusion between falling and sitting classes (a 10%). Furthermore, running class is confused with walking one the 37,5% times, an error that could be due to the small size of the training set.

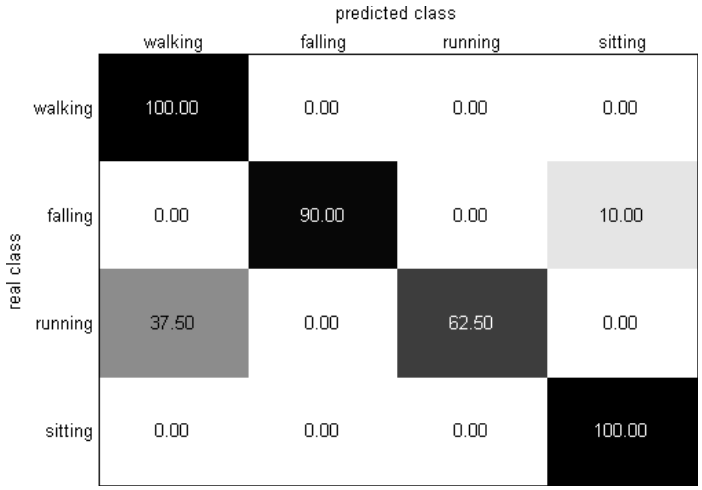

Fig. 7. Confusion matrix obtained, using only the new dataset generated for this work.

There is a little confusion between falling and sitting classes (a 10%). Furthermore, running class was confused with walking class the 37,5% times (but not the other way). This errors could be due the training set is too small.

6 Conclusions and Future Work

In this paper, an off-the-shelf solution for action recognition in video-surveillance applications from monocular RGB sequences has been proposed, implemented and successfully tested.

From the results shown in the paper, it can be concluded that the proposal is further than reliable, accurately discriminating among even the most similar activities. Besides, and thanks to the modeless solution adopted, the proposal shows high robustness, maintaining the reliability of the results within really different environmental situations, image point of view and scale.

Further from being this the main contribution presented in the paper, a dataset specifically designed for this application in the context of human activity recognition has been recorded and annotated, and is ready to be distributed among the scientific community, solving therefore this important detected need.

Nevertheless, it has been noticed that it is necessary to record a larger number of sequences to improve the training of SVM, and thus the classifier insensitivity and robustness.

As a future work, the modification of this proposal to be used with RGBD sequences is being tacked with still few but promising results.

References

1. D. A. Reid, S. Samangooei, C. Chen, M. S. Nixon, and A. Ross. Soft Biometrics for Surveillance: An Overview. In Handbook of statistics, 31(13), pp 327-351. Elsevier (2013).
2. B. E. Mennecke and A. Peters. From avatars to mavatars: The role of marketing avatars and embodied representations in consumer profiling. Business Horizons, 56(3), pp. 387-397, (2013).
3. A. Marcos-Ramiro, D. Pizarro, M. Marron-Romera and D. Gatica-Perez. Let Your Body Speak: Communicative Cue Extraction on Natural Interaction Using RGBD Data. In , IEEE Transactions on Multimedia, 17(10), pp. 17211732, (2015).
4. Dragan, M.A., Mocanu, I.; Human activity recognition in smart environments. 19th International Conference on Control Systems and Computer Science (CSCS) 2013, pp. 495-502 (2013).
5. Jalal, A., Uddin, M., Kim, T.S.: Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home. IEEE Transactions on Consumer Electronics, 58(3), 863-871. (2012).
6. Bengalur, M.: Human activity recognition using body pose features and support vector machine. Advances in Computing, 2013 International Conference on Communications and Informatics (ICACCI), pp. 1970-1975 (2013).
7. Ashraf, N., Sun, C., Foroosh, H.: View invariant action recognition using projective depth. Computer Vision and Image Understanding 123(0), 41-52 (2014).
8. Hu, N., Englebienne, G., Lou, Z., Krose, B.: Learning latent structure for activity recognition. 2014 IEEE International Conference on Robotics and Automation (ICRA), pp. 1048-1053 (2014).
9. Zhu, Y., Chen, W., Guo, G.: Evaluating spatiotemporal interest point features for depth-based action recognition. Image and Vision Computing 32(8), 453-464 (2014).
10. Han, J., Shao, L., Xu, D., Shotton, J.: Enhanced computer vision with Microsoft Kinect sensor: A review. IEEE Transactions on Cybernetics, 43(5), 1318-1334 (2013)
11. Kinect for Windows.- Microsoft. `https://www.microsoft.com/en-us/kinectforwindows/`
12. Sadanand, S., Corso, J.J.: Action bank: A high-level representation of activity in video. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2012).

13. El Houda Slimani, K., Benezeth, Y., Souami, F.: Human interaction recognition based on the co-occurrence of visual words. 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 461-466 (2014).
14. Hasan, M., Roy-Chowdhury, A.: Incremental activity modeling and recognition in streaming videos. 2014 IEEE Conference onComputer Vision and Pattern Recognition (CVPR), . pp. 796-803 (2014).
15. Oshin, O., Gilbert, A., Bowden, R.: Capturing relative motion and finding modes for action recognition in the wild. Computer Vision and Image Understanding 125(0), pp. 155-171 (2014).
16. Pehlivan, S., Forsyth, D.A.: Recognizing activities in multiple views with fusion of frame judgments. Image and Vision Computing 32(4), pp. 237-249 (2014).
17. Tran, K., Kakadiaris, I., Shah, S.: Part-based motion descriptor image for human action recognition. Pattern Recognition 45(7), pp. 2562-2572 (2012).
18. Bebars, A., Hemayed, E.: Comparative study for feature detectors in human activity recognition. 2013 9th International Computer Engineering Conference (ICENCO), . pp. 19-24 (2013).
19. Sanroma, G., Patino, L., Burghouts, G., Schutte, K., Ferryman, J.: A unified approach to the recognition of complex actions from sequences of zone-crossings. Image and Vision Computing 32(5), pp. 363-378 (2014).
20. Gall, J., Yao, A., Van Gool, L.: 2d action recognition serves 3d human pose estimation. In: Proceedings of the 11th European Conference on Computer Vision Conference on Computer Vision: Part III (ECCV'10). pp. 425-438, Springer-Verlag, Berlin, Heidelberg (2010).
21. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1297-1304. (2011).
22. Sapp, B., Weiss, D., Taskar, B.: Parsing human motion with stretchable models. Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1281-1288 (2011)
23. Pirsiavash, H., Ramanan, D.: Parsing videos of actions with segmental grammars. 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 612-619 (2014).
24. Alexandros André Chaaraoui, Pau Climent-Pérez, and Francisco Flórez-Revuelta. A review on vision techniques applied to human behaviour analysis for ambient-assisted living. Expert Systems with Applications, 39(12), pp. 1087310888 (2012).
25. CAVIAR Project (2004). Caviar test case scenarios. `http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/` (accesed: january 2016)
26. D. Weinland, R. Ronfard, E. Boyer. Free viewpoint action recognition using motion history volumes. Computer Vision and Image Understanding, 104, pp. 249257 (2006).
27. KTH - Recognition of human actions. `http://www.nada.kth.se/cvap/actions/`
28. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. Proceedings of the 17th International Conference on Pattern Recognition, (ICPR'04), 3, pp. 32-36. IEEE Computer Society, Washington, DC, USA (2004).
29. S. Singh, S. Velastin, H. Ragheb. Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. 2010 Seventh IEEE international conference on advanced video and signal based surveillance (AVSS), IEEE, pp. 4855 (2010).

30. Rodriguez, M., Ahmed, J., and Shah, M. (2008). Action mach a spatio-temporal maximum average correlation height filter for action recognition. IEEE conference on computer vision and pattern recognition, CVPR 2008 pp. 18 (2008).
31. L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri. Actions as space-time shapes. Transactions on Pattern Analysis and Machine Intelligence, 29 , pp. 22472253 (2007).
32. , Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: Proceedings of the British Machine Vision Conference (BMVA), pp. 124.1-124.11, (2009).
33. Klaeser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. Proceedings of the British Machine Vision Conference (BMVA), pp. 99.1-99.10 (2008).
34. Grandvalet, Y., Guigue, V., Rakotomamonjy, A., Canu, S.: SVM and Kernel Methods Matlab Toolbox. Perception Systemes et Information, INSA de Rouen, Rouen,France (2005)
35. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), pp. 273-297 (1995)
36. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, pp. 27:1-27:27 (2011).
37. C. Cortes and V. Vapnik. Support-vector network. Machine Learning, 20, pp. 273-297 (1995).
38. B. Scholkopf, A. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. Neural Computation, 12, pp. 1207-1245 (2000).