MEMOIRS OF THE SCIENTIFIC SECTIONS OF THE ROMANIAN ACADEMY

SECTION TITLE (E.G., MATHEMATICS, COMPUTER SCIENCE, BIOLOGY, CHEMISTRY HISTORY OF SCIENCE AND TECHNOLOGY)

HUMAN-ROBOT INTERACTION THROUGH ROBUST GAZE FOLLOWING

SORIN M. GRIGORESCU✉

FLORIN MOLDOVEANU

Department of Automation
Transilvania University of Brasov
Mihai Viteazu 5, 500174, Brasov, Romania[1]

In this paper, a probabilistic solution for gaze following in the context of joint attention will be presented. Gaze following, in the sense of continuously measuring (with a greater or a lesser degree of anticipation) the head pose and gaze direction of an interlocutor so as to determine his/her focus of attention, is important in several important areas of computer vision applications, such as the development of non-intrusive gaze-tracking equipment for psychophysical experiments in Neuroscience, specialized telecommunication devices, *Human-Computer Interfaces* (HCI) and artificial cognitive systems for *Human-Robot Interaction* (HRI). We have developed a solution based on a probabilistic approach that inherently deals with uncertainty of sensor models and incomplete data. This solution comprises a hierarchical formulation of a set of detection classifiers that loosely follows how geometrical cues provided by facial features are used by the human perceptual system for gaze estimation. A quantitative analysis of the proposed architecture's performance was undertaken through a set of experimental sessions. In these sessions, temporal sequences of moving human agents fixating a well-known point in space were grabbed by the stereovision setup of a robotic perception system, and then processed by the framework.

Keywords: Gaze estimation, Feedback control in image processing, Facial features detection, Human-Robot Interaction.

1. INTRODUCTION

Head movements are commonly interpreted as a vehicle of interpersonal communication. For example, in daily life, human-beings observe head movements as the

[1] Emails: s.grigorescu@unitbv.ro, moldof@unitbv.ro

expression of agreement or disagreement in a conversation, or even as a sign of confusion. On the other hand, gaze shifts are usually an indication of intent, as they commonly precede action by redirecting the sensorimotor resources to be used. As a consequence, sudden changes in gaze direction can express alarm or surprise. Gaze direction can also be used for directing a person to observe a specific location. To this end, during their infancy, humans develop the social skill of *joint attention*, which is the means by which an agent looks at where its interlocutor is looking at by producing an eye-head movement that attempts to yield the same focus of attention. Over nine months of age, infants are known to begin to engage with their parents/caregivers in an activity in which both look at the same target through joint attention.

As artificial cognitive systems with social capabilities become more and more important due to the recent evolution of robotics towards applications where complex and human-like interactions are needed, basic social behaviors such as joint attention have increasingly become important research topics in this field. Fig. 1 illustrates the ROVIS[2] (*Robust Vision and Control Laboratory*) gaze following system at work, under the context of joint attention for Human Robotic Interaction (HRI). Gaze following thus represents an important part of building a social bridge between humans and computers. Researchers in robotics and artificial intelligence have been attempting to accurately reproduce this type of interaction in the last couple of decades, and, although much progress has been made [1], dealing with perceptual uncertainty still renders it difficult for these solutions to work adaptively.

[2] http://rovis.unitbv.ro

Fig. 1 - Gaze following in the context of joint attention for HRI, using the ROVIS system on a Neobotix MP 500 mobile platform.

Gaze following is an example for which the performance of artificial systems are still far from human adaptivity. In fact, the gaze following adaptivity problem can be stated as follows: how can gaze following be implemented under non-ideal circumstances (perceptual uncertainty, incomplete data, dynamic scenes, etc.)? Fig. 2 demonstrates how incomplete data, arguably the issue where the lack of adaptivity and underperformance of artificial systems are most apparent, might influence the outcome of gaze following.

Fig. 2 - Examples of probable gaze following failure scenarios due to incomplete data: facial features occluded in profile views (a), or failure of feature detection algorithms (b).

Feature detection represents a subtopic within the head pose estimation problem. Accurate estimate for the eye, nose or the mouth represents an intermediate stage, in

which essential information used by the geometrical approach for head pose estimation is computed. Mouth recognition is dealt with methods such as the ones suggested in [2] and [3]. A common approach for detecting the mouth is by pre-segmenting red color on a specific patch of the image. Both methods use a ROI extracted after head segmentation, in which the mouth is approximately segmented, after a color space conversion is performed (such as RGB to HSI (*Hue, Saturation, Intensity*) [2], or RGB to Lab [3]). On the other hand, nose detection algorithms use Boosting classifiers, commonly trained with Haar-like features [4], or the 3D information of the face, as in [5]. The shape-based algorithm proposed in [6], built on the isophote curvature concept, i.e. the curve that connects points of the same intensity, is able to deliver accurate eye localization from a web camera. The eye location can be determined using a combination of Haar features [7], dual orientation Gabor filters and eye templates, as described in [8].

In the following text, we propose a robust solution to facial feature detection for human-robot interaction based on a i) feedback control system implemented at the image processing level for the automatic adaptation of the system's parameters, ii) a cascade of facial features classifiers and iii) a *Gaussian Mixture Model* (GMM) for facial points segmentation. The goal is to obtain a real-time gaze following estimator capable of dealing with perceptual uncertainty and incomplete data. The expected outcome of this project will be an autonomous system, with the ability of robustly estimating the gaze's direction of interlocutors within the context of joint attention in HRI.

2. CONTROLLING A MACHINE VISION SYSTEM

In a robotic application, industrial or real world, the purpose of the image processing system is to understand the surrounding environment of the robot through visual information.

Low level image processing deals with pixel wise operations aiming to improve the input images and also separate the objects of interest from the background. Both the inputs and outputs of low level blocks are images. The second type of modules, which deal with high level visual information, are connected to low level operations through the feature extraction component which converts the input images to abstract data describing the imaged objects of interest. For the rest of the high level operations, both the inputs and outputs are abstract data. The importance of the quality of the results coming from low

level stages is related to the requirements of high level image processing [9]. Namely, in order to obtain a proper visual understanding of the imaged environment at high level stage, the inputs coming from low level have to be reliable.

The sequential, feedback free, approach has an impact on the final perception result, since each operation in the chain is applied sequentially, with no information between the different levels of processing. In other words, low level image processing is done regardless of the requirements of higher levels. For example, if the segmentation module fails to provide a good output, all the subsequent steps will fail. In [10] and [11], the inclusion of feedback structures within vision algorithms for improving the overall robustness of the chain is suggested. In the proposed approach, the parameters of low level image processing are adapted in a closed-loop manner in order to provide reliable input data to higher levels of processing.

The basic diagram, from which the feedback mechanisms for machine vision are derived in this paper, can be seen in Fig. 3. In such a control system, the control signal u, or *actuator variable*, is an image processing parameter, whereas the *controlled variable* y is a measure of feature extraction quality.

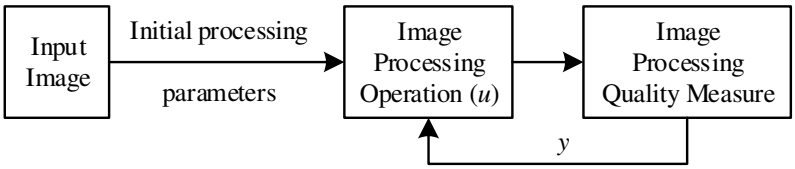

Fig. 3 - Feedback adaptation of an image processing operation. The image processing quality measure y is used as a feedback control variable for adapting the parameters of the vision algorithms using the actuator u.

3. IMAGE PROCESSING CHAIN

The gaze following image processing chain, depicted in Fig. 4, contains four main steps. We assume that the input is an 8-bit gray-scale image $I = J^{V \times W}$, of width V and height W, containing a face viewed either from a frontal or profile direction, where $J = \{0, \dots, 255\}$. (v, w) represents the 2D coordinates of a specific pixel. The face region is obtained from a face detector.

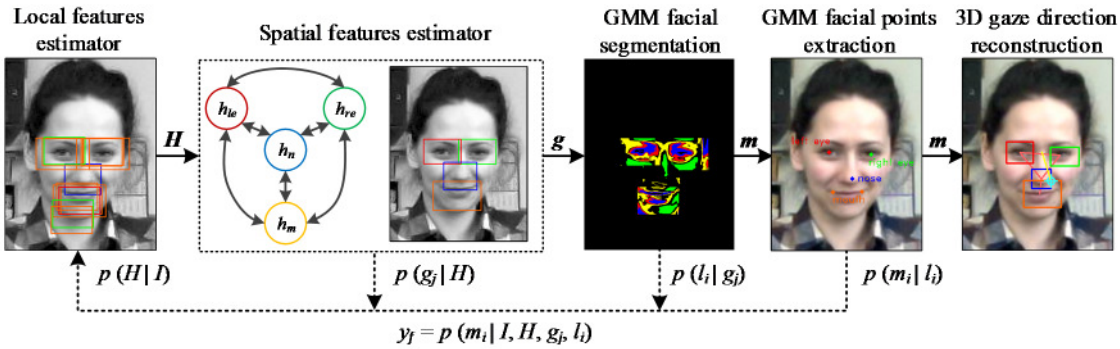

Fig. 4 - Block diagram of the proposed gaze following system for facial feature extraction and 3D gaze orientation reconstruction. Each processing block within the cascade provides a measure of feature extraction quality, fused within the controlled variable y_f (see Eq. 2).

Firstly, a set of facial features ROI hypotheses $H \in \{h_{le}, h_{re}, h_n, h_m\}$, consisting of possible instances of the left h_{le} and right h_{re} eyes, nose h_n and mouth h_m, are extracted using a local features estimator which determines the probability measure $p(H|I)$ of finding one of the searched local facial region. The number of computed ROI hypotheses is governed by a probability threshold T_h, which rejects hypotheses with a low $p(H|I)$ confidence measure. The choice of the T_h threshold is not a trivial task when considering time critical systems, such as the gaze estimator, which, for a successful HRI, has to deliver in real-time the 3D gaze orientation of the human subject. The lower T_h is, the higher the computation time. On the other hand, an increased value for T_h would reject possible "true positive" facial regions, thus leading to a failure in gaze estimation. As explained in the followings, in order to obtain a robust value for the hypotheses selection threshold, we have chosen to adapt T_h with respect to the confidences provided by the subsequent estimators from Fig. 4, which take as input the facial regions hypotheses. The output probabilities coming from these estimation techniques, that is, the spatial estimator and the GMM for pointwise feature extraction, are used in a feedback manner within the extremum seeking control paradigm.

Once the hypotheses vector H has been build, the facial features are combined into the spatial hypotheses $g = g_0, g_1, \ldots, g_n$, thus forming different facial regions combinations. Since one of the main objective of the presented algorithm is to identify facial points of frontal, as well as profile faces, a spatial vector s_i is composed either from four, or three, facial ROIs:

$$g_i = \{h_0, h_1, h_2, h_3\} \cap \{h_0, h_1, h_2\} \quad (1)$$

where $h_i \in \{h_0, h_1, h_2, h_3\}$.

The extraction of the best spatial features combination can be seen as a graph search problem $g_j = f: G(g, \mathbf{E}) \rightarrow \mathbb{R}$, where $\mathbf{E}$ are the edges of the graph connecting the hypotheses in g. The considered features combinations are illustrated in Fig. 5. Each combination has a specific spatial probability value $p(g_j|H)$ given by a spatial estimator trained using the spatial distances between the facial features from a training database.

Once the spatial distributions of the probable locations of the facial features ROIs are available, their point-wise location m_i is determined using a GMM segmentation method. Its goal is to extract the most probable facial point-wise locations m_i given the GMM pixel likelihood values $p(l_i|g_j)$. The most relevant point features for computing the 3D gaze of a person are the centers of the eyes, tip of the nose and corners of the mouth. The described data analysis methods are used to evaluate a feature space composed of the local and spatial features.

Having in mind the facial feature points extraction algorithm described above, it can be stated that the confidence value y_f of the processing chain in Fig. 4 is a probability confidence measure obtained from the estimators cascade:

$$y_f = p(m_i|I, H, g_j, l_i) \tag{2}$$

Since the whole described processing chain is governed by a set of parameters, such as the threshold T_h for selecting the vector s, we have chosen to adapt it using an extremum seeking control mechanism and the feedback variable y_f, derived from the output of the gaze following structure illustrated in Fig. 4. The final 3D gaze orientation vector $\vec{\varphi}(m_i)$, representing the roll, pitch and yaw of the human subject, is determined using the algorithm proposed in the work of Gee and Cipolla [12].

4. PERFORMANCE EVALUATION

4.1. EXPERIMENTAL SETUP

In order to test the performance of proposed gaze following system, the following experimental setup has been prepared.

The system has been evaluated on the *Labeled Faces in the Wild* (LFW) database [13]. LFW consists of 13.233 images, each having a size of $250 \times 250px$. In addition to the LFW database, the system has been evaluated on an Adept Pioneer 3-DX mobile robot equipped with an RGB-D sensor delivering $640 \times 480px$ size color and depth images. The goal of the scenarios is to track the facial features of the human subject in the HRI context. The error between the real and estimated facial feature's locations was computed offline.

For evaluation purposes, two metrics have been used:

- the mean normalized deviation between the ground truth and the estimated positions of the facial features:

$$d(\mathbf{m}, \hat{\mathbf{m}}) = \tau(\mathbf{m}) \frac{1}{k} \sum_{i=0}^{k-1} \|m_i - \hat{m}_i\| \tag{3}$$

where k is the number of facial features, $\mathbf{m}$ and $\hat{\mathbf{m}}$ are the manually and estimated annotated positions of the eyes, nose and mouth, respectively, and $\tau(\mathbf{m})$ is a normalization constant:

$$\tau(\mathbf{m}) = \frac{1}{\|(m_{le} + m_{re}) - m_m\|} \tag{4}$$

- the maximal normalized deviation:

$$d^{\max}(\mathbf{m}, \hat{\mathbf{m}}) = \tau(\mathbf{m}) \max_{j=0,\ldots,k-1} \|m_i - \hat{m}_i\| \tag{5}$$

4.2. COMPETING DETECTORS

The proposed gaze following system has been tested against three open-source detectors.

1) *Independent facial feature extraction*: The detector is based on the Viola-Jones boosting cascades and returns the best detected facial features, independent of their spatial relation. The point features have been considered to be the centers of the computed ROIs.

The boosting cascades, one for each facial feature, has been trained using a few hundred samples for each eye, nose and mouth. The searching has been performed several times at different scales, with Haar-like features used as inputs to the basic classifiers within the cascade. From the available ROI hypotheses, the one having the maximum confidence value has been selected as the final facial feature.

2) *Active Shape Models*: An *Active Shape Model* (ASM) estimates a dense set of feature points distributed around face contours such as eyes, nose, mouth, eyebrows, or chin. An ASM is initially trained using a set of manually marked contour points.

The open-source AsmLib, based on OpenCV, has been used as candidate detector. The ASM is trained from manually drawn face contours. The trained ASM model calculates the main variations in the training dataset using *Principal Component Analysis* (PCA), which enables the model to automatically recognize if a contour is a face contour. PCA is used to find the mean shape and the main variations of the training data with respect to the mean shape. After finding the shape model, all training objects are deformed to the main shape, and the pixels converted to vectors. The positions of the contours at each search step are corrected by the usage of the lines perpendicular to the control points of the contour. After creating the ASM model, an initial contour is deformed by finding the best texture match for the control points. This is an iterative process, in which the movement of the control points is limited by what the ASM model recognizes from the training data as a ”normal” face contour.

3) *Flandmark*: *Flandmark* [14] is a deformable part model detector of facial features, where the detection of the point features is treated as an instance of structured output classification. The algorithm is based on a *Structured Output Support Vector Machine* (SO-SVM) classifier for the supervised learning of the parameters for facial points

detection from examples. The objective function of the learning algorithm is directly related to the performance of the resulting detector, which is controlled by a user-defined loss function.

In comparison to our gaze following system, which uses a segmentation step for determining the pointwise location of the facial features, Flandmark considers the centers of the detected ROIs as the point location of the eyes, nose, and mouth.

The mean and maximal deviation metrics were used to compare the accuracy of the four tested detectors with respect to the ground truth values available from the benchmark databases. Especially for the evaluation of the computation time, the algorithm has also been tested on a mobile robotic platform.

The cumulative histograms of the mean and maximal normalized deviation are shown in Fig. 6 for frontal and profile faces. In all cases, the proposed estimator delivered an accuracy value superior to the ones given by the competing detectors. If the accuracy difference between our algorithm and Flandmark is relatively low for the case of frontal faces, it actually increases when the person's face is imaged from a profile view.

An interesting observation can be made when comparing the independent detectors with the ASM one. Although the ASM outperforms independent facial feature extraction on frontal faces, it does not performs well when the human subjects are viewed from lateral. This is due to the training nature of the ASM, where the input training data is made of points spread on the whole frontal area (e.g. eyes, eyebrows, nose, chin, cheeks, etc.).

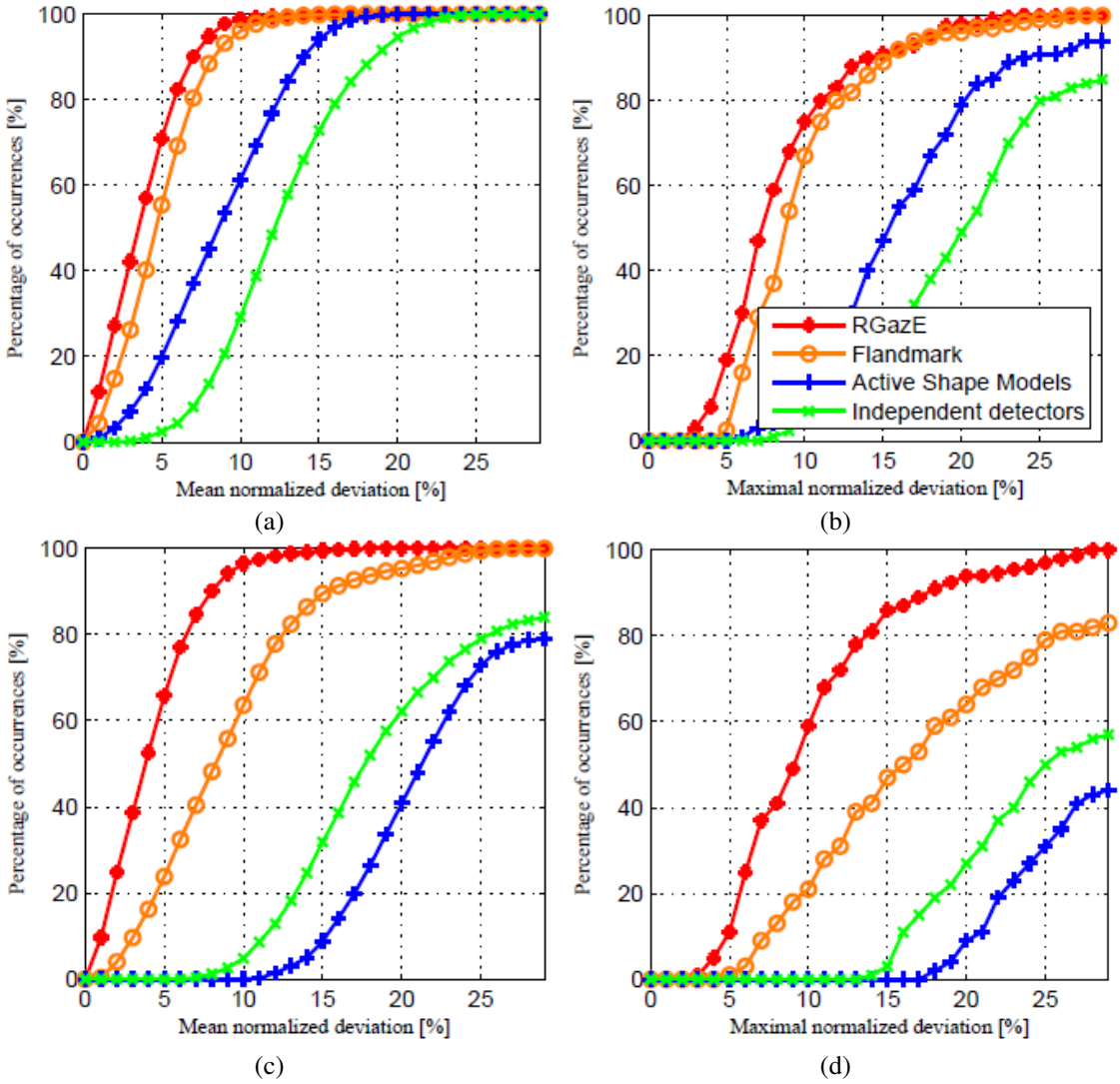

Fig. 5 - Cumulative histograms for the mean and the maximal normalized deviation shown for all competing detectors applied on video sequences with frontal (a,b) and profile (c,d) faces.

5. CONCLUSIONS

In this paper, a robust facial features detector for 3D gaze orientation estimation has been proposed. The solution is able to return a reliable gaze estimate, even if only a partial set of features is available, with a clear indication of the uncertainty involved. The paper brings together algorithms for facial feature detection, machine learning and control theory. During the experiments, we have investigated the system response and compared the results to ground truth values. As shown in the experimental results section, the method performed well with respect to various testing scenarios. As future work, the

authors consider the possibility of extending the framework for the simultaneous gaze estimation of multiple interlocutors and the adaptation of algorithm with respect to the robot's egomotion.

Acknowledgment. We hereby aknowledge the structural founds project PRO-DD (POS-CCE, O.2.2.1., ID 123, SMIS 2637, ctr. No 11/2009) for providing the infrastructure used in this work.

REFERENCES

[1] B. SCASSELLATI, *Theory of mind for a humanoid robot*, Autonomous Robots, vol. 12, no. 1999, pp. 13–24, 2002.

[2] M. PANTIC, M. TOMC, and L. ROTHKRANTZ, *A hybrid approach to mouth features detection*, in 2001 IEEE International Conference on Systems, Man, and Cybernetics, vol. 2, pp. 1188 –1193, 2001.

[3] E. SKODRAS and N. FAKOTAKIS, *An unconstrained method for lip detection in color images*, in 2011 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1013 –1016, 2011.

[4] D. GONZALEZ-ORTEGA, F. DIAZ-PERNAS, M. MARTINEZ-ZARZUELA, M. ANTON-RODRIGUEZ, J. DIEZ-HIGUERA, and D. BOTO-GIRALDA, *Real-time nose detection and tracking based on ADABoost and optical flow algorithms*, in Intelligent Data Engineering and Automated Learning. Springer, Berlin, vol. 5788, pp. 142–150, 2009.

[5] N. WERGHI, H. BOUKADIA, Y. MEGUEBLI, and H. BHASKAR, *Nose detection and face extraction from 3D raw facial surface based on mesh quality assessment*, in 36th Annual Conference on IEEE Industrial Electronics Society, pp. 1161 –1166, 2010.

[6] R. VALENTI, N. SEBE, and T. GEVERS, *Combining head pose and eye location information for gaze estimation*, IEEE Transaction on Image Processing, 2011.

[7] P. VIOLA and M. JONES, *Rapid object detection using a boosted cascade of simple features*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, 2001.

[8] L. KE and J. KANG, *Eye location method based on Haar features*, in 2010 3rd International Congress on Image and Signal Processing, vol. 2, pp. 925 –929, 2010.

[9] L. HOTZ, B. NEUMANN, and K. TERZIC, *High-level expectations for low-level image processing*, in KI 2008: Advances in Artificial Intelligence. Springer-Verlag Berlin Heidelberg, 2008.

[10] D. RISTIC, *Feedback structures in image processing*, Ph.D. dissertation, Bremen University, Institute of Automation, Bremen, Germany, Apr. 2007.

[11] S. M. GRIGORESCU, *Robust machine vision for service robotics*, Ph.D. dissertation, Bremen University, Institute of Automation, Bremen, Germany, June 2010.

[12] A. GEE and R. CIPOLLA, *Determining the gaze of faces in images*, Image and Vision Computing, vol. 12, no. 10, pp. 639–647, 1994.

[13] G. B. HUANG, M. RAMESH, T. BERG, and E. LEARNED-MILLER, *Labeled faces in the wild: A database for studying face recognition in unconstrained environments*, University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[14] M. URICAR, V. FRANC, and V. HLAVAC, *Detector of facial landmarks learned by the structured output SVM*, in VISAPP '12: Proceedings of the 7th International Conference on Computer Vision Theory and Applications, G. Csurka and J. Braz, Eds., vol. 1. Portugal: SciTePress — Science and Technology Publications, pp. 547–556, February 2012.