Detecting Conversational Gaze Aversion Using Unsupervised Learning

Matthew Roddy, Naomi Harte
ADAPT Centre, School of Engineering
Trinity College Dublin, Ireland

***Abstract*—The aversion of gaze during dyadic conversations is a social signal that contains information relevant to the detection of interest, turn-taking cues, and conversational engagement. The understanding and modeling of such behavior has implications for the design of embodied conversational agents, as well as computational approaches to conversational analysis. Recent approaches to extracting gaze directions from monocular camera footage have achieved accurate results. We investigate ways of processing the extracted gaze signals from videos to perform gaze aversion detection. We present novel approaches that are based on unsupervised classification using spectral clustering as well as optimization methods. Three approaches that vary in their input parameters and their complexity are proposed and evaluated.**

I. Introduction

Embodied conversational agents (ECAs), such as conversational robots or virtual avatars, need to be able to extract information about the attentional states of interlocutors to carry out naturalistic interactions. This attentional information is used to create engagement models that can control conversational decisions, such as when to begin and end interactions, turn-taking behavior, and what content is used. Errors in the engagement model can lead to interruptions, unanswered questions, false starts, and false endings. These models can be created using a variety of features from different modalities (e.g. linguistic, prosodic, visual). Gaze-based features have been established to be indicators of attention and interest [1], [2]. Additionally, gaze aversion has been linked to emotions such as embarrassment and shame [3], as well as disagreement [4]. As such, automatic methods of detecting gaze aversion are highly relevant to the implementation of ECAs that can respond to, and mimic, human social signals.

Up until recently, conversational gaze analysis required the use of either specialized equipment (eye-tracking glasses, infrared cameras), camera calibration stages, or laborious hand-annotation. Recent advances in appearance-based gaze tracking [5], [6] have facilitated the extraction of gaze vectors and head pose information from standard monocular video signals. We propose a novel application of unsupervised learning for detecting conversational gaze aversion that relies on the extracted gaze vectors and head pose measurements. Our approach abstracts the gaze aversion classification problem from gaze-tracking and pose-estimation procedures. Consequently, the overall performance of our approach should increase as the accuracy of gaze trackers increases.

Our approach is based on using the gaze vectors and pose signals extracted from videos of a dyadic conversation to create two separable clusters of gaze vector projections that correspond to when a gazer is looking at the target interlocutor's head, and when the gazer is averting their gaze. We refer to the person whose gaze aversions we are trying to detect as the "gazer", and the person being looked at as the "target". Our aim is to make the clusters as separable as possible. We present three different methods of creating separable clusters that vary in increasing degrees of complexity. We then use spectral clustering (SC) [7] to classify the clusters in an unsupervised manner.

In II we present an overview of previous work on gaze aversion, gaze tracking and relevant aspects from psychology. In III we present an overview of our approach, discussing aspects of SC that are important for our classification methods. In IV the feature extraction system is discussed, as well as the data set. In V we discuss the details of the three different methods and show their performance on the data set. In VI we discuss the results and present our observations on the performance of the three methods. Finally, in VII we summarize and present our intentions for how this work will be applied and advanced.

II. Previous work on gaze aversion

Gaze aversion has been shown to play a role in turn-management [8] and is also used as a way of limiting cognitive load whilst considering responses to questions [9], [10]. These "thinking" gaze gestures are saccades in the direction of uninformative regions of space that are often subtle and difficult even for humans to identify [11]. Algorithms that can detect turn-management gaze gestures and "thinking" gaze aversion gestures are needed for the design of naturalistic ECAs.

An area that has received a large amount of recent interest is gaze-tracking using appearance-based models [12], [6], [5], [13]. Appearance-based gaze tracking enables gaze gesture analysis by performing gaze tracking on standard monocular video signals. However, the classification problem of conversational gaze aversion detection has received less attention. In [11], Morency *et al.* proposed an SVM-based approach that discriminates "thinking" gaze aversion gestures from eye contact and deictic gestures (saccades that reference a specific object or person of interest) using temporal windows of eye gaze directions. In [14] the SVM-based approach is improved upon with the use of Latent-Dynamic Conditional Random Fields (LDCRFs). Methods of deducing visual focus of attention (VFOA) [15], [16], [17] could also be used to infer

ISBN 978-0-9928626-7-1

gaze aversion. Many VFOA methods rely on head orientation estimation to distinguish the focus of attention in multi-party meeting scenarios. While these methods could be used to infer gaze aversion gestures, saccades that constitute "thinking" eye gestures may be more difficult to detect due to their subtle nature.

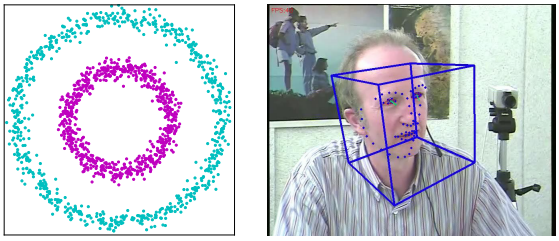
III. Unsupervised gaze analysis

A. Overview

Our approach to conversational gaze-aversion detection is based on the observation that during dyadic conversation, when a gazer is looking at a target interlocutor, the gaze vectors will cluster in the direction of the target's head. When the gazer averts their gaze, the vectors will point anywhere in the space around the target's head. If the gaze aversions occur in many directions around the target's head, two clusters are created, an inner cluster and an outer circular cluster. As such, our problem is a two-class clustering problem. Spectral clustering (discussed below in III-B) is well suited to this manner of classification problem. We present three gaze aversion detection methods that use SC as their core classification algorithm and evaluate them in three experiments. The three methods are designed to increase the separability of the two clusters and they vary incrementally in their complexity.

If the gaze vectors are noisy, the definition between the clusters is blurred. We present a simple method to address this in the first experiment. Secondly, if the pose of the gazer changes over the course of the conversation (assuming the target is static), the gaze directions for the two classifications will overlap. We address this problem in the second experiment. Thirdly, if the target's pose changes over the course of the conversation, the gaze classifications will overlap. We address this problem in the second and third methods. Additionally, in the third method, we propose a way of incorporating rough estimates of relative camera distances and angles, if they can be inferred.

B. Spectral Clustering



(a) (b)

Fig. 1. a) Example of SC performance on a nested circle classification problem. b) Example of extracted gaze vectors and head pose.

Spectral clustering [18] is known to perform well in clustering applications where other methods (naive k-means and mean-shift) fail. Spectral clustering approximates solutions to normalized graph cuts. It has been shown to perform well on the on the problem of clustering nested circles [19] as shown in Fig. 1a. We use an RBF kernel to when calculating the affinity matrix to make the inner nested cluster more separable. The RBF Kernel is given by $K(x, x') = \exp(-\gamma\|x - x'\|^2)$ where $\gamma = \frac{1}{2\sigma^2}$. The value of the RBF kernel decreases as the squared euclidean distance increases. This causes points that lie on a circle around a given point x to have the same kernel value. The γ value can be intuitively thought of as a parameter that controls the distance of influence of a given point. A small γ indicates a high variance and a large region of influence.

In the current application, the γ value can be thought of as a parameter that controls the variance of the central 'gaze-on' cluster. Higher variances imply clusters that are more spread out. This optimal value for this parameter in our application is therefore dependent on how far away the person is from the camera how large the gaze-away angles are. While the optimal value for this parameter is difficult to infer precisely for a given recording, for our experiments we found the spectral clustering algorithm to still perform well under a variety of γ values. In our experiments we use γ values that vary between 0.1 and 1.0 in increments of 0.1. We then selected the γ value that yielded the best performance.

IV. Experimental setup

A. Data set and extracted features

We use OpenFace [5] to extract features for speakers in 9 dyadic conversations selected from the IFA Dialog Video Corpus [20]. The features include: gaze unit vectors for both eyes, eye locations, head pose, head rotation, blink times, and feature confidence scores. An example of a frame with extracted gaze vectors and head pose measurements is shown in Fig. 1b. All vector, distance, and rotation measurements were in 3-dimensions. Each conversation in the data set uses a two-camera setup, where each camera is synchronized with the other at a frame rate of 25 frames/sec, and each conversation lasts 15 minutes. Each conversation includes binary gaze annotations for when each person is looking at the other interlocutor. We use these annotations as the ground-truth against which our gaze-aversion classifier is tested.

Figure 2 shows a diagram of the camera setup as seen from above. The two interlocutors were seated across from each other at a table and the cameras were placed behind each speaker's left shoulder to capture the face of the person across the table. The relative angles and distances between the cameras are different for each recording because the camera setup had to be dismantled after each recording session. The focal lengths of each of the cameras were also adjusted for each of the conversations. These distances, angles and focal lengths were not recorded during the sessions.[1] However, the lens models for each camera were recorded. We used these to limit the possible focal length values. It is also noted in [20] that the speakers were seated roughly one meter across from each other at a table. Using focal length estimates, the distance estimate, and the gaze annotations, we devise a system to

[1]This was verified through private correspondence with one of the authors of [20]

ISBN 978-0-9928626-7-1 © EURASIP 2017

estimate the relative position and rotations of the cameras in section V-C.

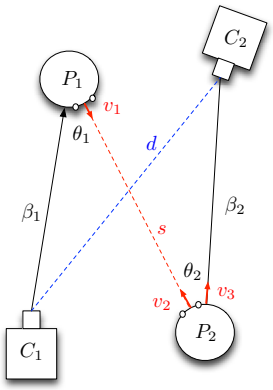


Fig. 2. Camera setup

The extracted features from the video files were filtered on the basis of the confidence score output by OpenFace. This was necessary to deal with frames where the face was partially or totally covered, or where the eyes were not visible due to head rotation. Frames below the confidence threshold of 0.95 were removed from both the gazer and target features. We also filtered out blinks using the blink detection built into OpenFace. After removing these frames, the number of frames that we used to test our classifiers was 168,400 (112.3 minutes).

The main metric that we use to evaluate our algorithms is the Adjusted Rand Index (ARI), which computes a similarity measure between the labelings of two partitions of a set by performing pairwise comparisons [21]. A returned value of zero represents a performance equivalent to chance, and a value of 1 represents perfect labeling. In evaluating binary clusters this ARI metric is preferable to accuracy or F-scores as it takes into account the sizes of the clusters, as well as evaluates the performance on both classification labelings. We include accuracy and F-scores in Table I as reference. We performed McNemar tests on the results of the different methods (e.g method 1 vs method 2) to test the significance of the differences between the methods. In all cases the results were highly significant ($p \ll 0.001$).

V. Three approaches to gaze aversion detection

TABLE I
EXPERIMENTAL RESULTS FROM THE THREE METHODS AND THEIR DIFFERENT SETTINGS.

Method: settings	ARI	Accuracy (%)	F-score
1: Eye 1	0.236	78.1	0.478
1: Eye 2	0.252	79.0	0.489
1: Eyes 1+2	0.272	79.9	0.506
2: w/ gazer pose	0.363	84.1	0.565
2: w/ gazer/target pose	0.358	83.9	0.561
3: no target pose	0.335	82.4	0.554
3: w/ target pose	0.39	84.8	0.589

A. First method: Only gaze vectors

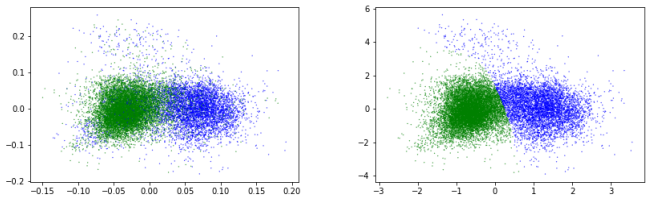


Fig. 3. The performance of SC on the x,y components of the gaze vectors. The true classes are shown on the left and the classifier performance is shown on the right.

The first method is our baseline system for gaze aversion classification. Firstly, we center and scale the gaze vectors for both eyes by subtracting the mean and dividing by the standard deviation (z-scores). We then perform SC on the x and y values of the vectors. The results are shown in the first two rows of Table I. In the table, accuracy and ARI includes classification of 'gaze-on' and 'gaze-away'. F-score is given in terms of how well the system classifies 'gaze-away'. We find that the right eye of the gazer (eye two) performs consistently better than eye one. This is due to the right eye being closer to the camera. Based on the observation that summing two vectors that point the same direction together will create a larger vector in that direction, we experiment with summing the two eye vectors together. We find that this improves performance when compared with the best performing single eye. Consequently, on the subsequent experiments we only report results on the combined eye performance.

B. Second method: Gaze vectors and gazer/target pose

In our second method we address the influence of both the gazer and the target's head pose. In Fig. 4a, the x-component of the gaze vectors is plotted against time, with the true classifications denoted by the colors. The plot shows that there is a shift in the trend of the x-component over the course of the conversation. This is caused by either the target, the gazer, or a combination of both, shifting their heads. We first try to remove the effect of the gazer's head movement.

We preprocess the pose signals by smoothing the pose locations using a five-point moving-average filter to reduce the influence of noise. We then attempt to estimate the amount of each pose dimension that can be removed from the gaze vectors by solving the linear least squares problem:

$$\min_{L} \|P \cdot L - D\|_2^2 \tag{1}$$

where L is a 2×2 matrix, P is an $N \times 2$ matrix of the smoothed gazer pose values, D is an $N \times 2$ matrix of the gaze vectors and N is the number of frames. The resulting value of $P \cdot L - D$ gives an estimate of the gaze vectors with the gazer pose component removed. In Fig. 4b the effect of removing the gazer's pose component is visible by the reduction in the the curved trend. In the fifth row of Table I we see that the

ISBN 978-0-9928626-7-1 © EURASIP 2017

overall performance on the data set is substantially increased by removing the gazer pose.

We try to further improve on these results by fitting a similar least-squares model of the target's pose to the gaze vectors that have the gazer's pose removed. This is motivated by the observation that if the target moves their head around, gaze vectors that point in the direction of a gaze aversion and gaze vectors that point in the direction of the target could overlap. The results in the sixth row of Table I show that this reduces the performance of the classifier. There are a number of potential reasons for this. The main one is that the angle formed by the directions of the two cameras (see Fig. 2) creates a nonlinear relationship between the pose values extracted in one camera system, and the pose values extracted in the other. We attempt to address this problem in the third method.

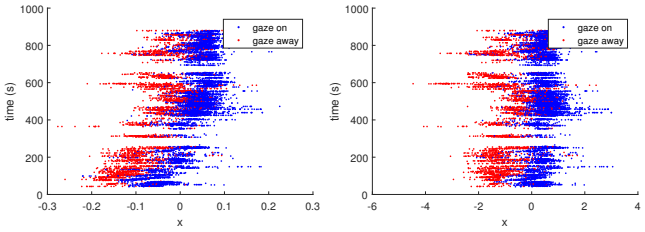


Fig. 4. The effect of removing the gazer's head movement component

C. Third method: Gaze vectors, gazer/target pose and camera position estimates

In this third method we attempt to estimate the relative angles and distances between the two cameras using the geometry of the camera setup and the gaze annotations. We then project the gaze vectors onto an affine hyperplane that is parallel to the target's head pose. This is an effort to simulate a situation where the rotations and distances were recorded when the data set was created. To estimate the relative rotation and distance, we approach the problem from a multi-sensor fusion perspective where two coordinate systems need to be fused. We treat the target's camera as the baseline coordinate system and project the gazer's camera measurements into this coordinate system. We use non-linear optimization on the compound manifold $SO(2)\times\mathbb{R}^3$, where the operator '$\times$' is the Cartesian product between two sets, and $SO(2)$ is the special orthogonal group in two dimensions. We optimize for both the angle and the distance simultaneously.

The system is founded on three assumptions: (1) that when mutual gaze occurs (when both people are looking at each other), gaze vectors for each person's eyes project onto the corresponding eye locations of the other person, (2) the average distance between the two people is 1 meter, (3) that the rotation of camera 2 is restricted to rotations around the y-axis. This third assumption is equivalent to assuming that the camera angles are parallel to the ground. From an informal analysis of the videos this assumption appears to be reasonable. We define an affine hyperplane in two dimensions that both target eyes are projected onto. The euclidean distance between the projection of the target eyes and the corresponding projection from the gaze vectors is the distance we minimize. The problem of finding the optimal distance and rotation values is then formulated as a non-linear least squares problem: $\arg\min \frac{1}{2}\|F(X)\|^2$. The full derivation of the cost function $F(X)$ is given online for the reader, as space does not allow us to reproduce it here.[2] We use the Levenberg-Marquardt algorithm as a solver and the Manifold Toolkit for Matlab (MTKM) [22] to simplify the manifold projections.

Once estimates for the distance and rotation values are found we project the gaze vectors onto the hyperplanes defined by the target head pose and perform SC as in the previous two methods. The penultimate row of Table I shows that the performance of the third method, as outlined above, is better than the first method but worse than the second method. We try to improve upon these results and find that removing the target pose component from the gaze projections in the same manner as Eq. 1 improves the results. The final line of Table I shows that we achieve better results with this than experiment two.

VI. Discussion

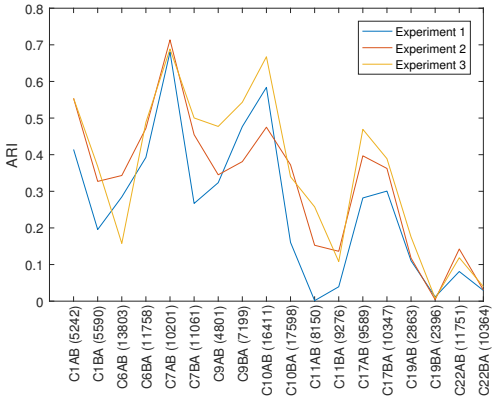


Fig. 5. Performance of the different experiments. The x-axis gives the test configuration ('C', conversation, gazer, target) followed by the number of frames (after filtering blinks and low confidence frames) that are used in the classification.

The variability of each method's performance on the different files is shown in Fig. 5. There are a number of potential explanations for the range of results. Firstly, if the vector angles created by the gaze aversion gestures are not large enough, the two clusters will be difficult to distinguish. The best performance (file set "C7AB") has clearly separable clusters where the gaze aversion angles are large. Secondly, if the gazer does not avert their gaze enough, the gaze aversion cluster will be difficult for the SC algorithm to identify. This behavior is what causes the poor performance in file set "C11BA". The performance of our presented approach is therefore dependent on the gaze behavior of the conversation participants. In file set

[2]www.github.com/mattroddy/eusipco_derivation

ISBN 978-0-9928626-7-1 © EURASIP 2017

"C19BA" we observe that the gazer's head rotations, relative to the camera, are extreme in comparison to those of the rest of the data set. We surmise that gaze measurement performance of OpenFace is worse when there are large head rotation angles. Our approach could potentially be refined to account for some of these factors.

During the first experiment we performed an informal test of different focal length parameters. We tested the maximum and minimum settings of the lens focal lengths. We found that after the scaling operation the resulting x,y values were identical for all settings. A consequence of this is that the first and second methods can be performed on any video file without the knowledge of the focal length settings. The relative distances between the clusters will be the same once they are scaled. This allows for the use of these algorithms in a wide variety of applications. Practical applications of our unsupervised approach to gaze aversion detection could be: large scale analysis of online videos (e.g YouTube videos), automatic analysis of audience interest, and automatic labeling of conversational data sets.

VII. Conclusions

We have presented a framework for detecting conversational gaze aversion from gaze vectors and pose measurements that uses SC as its core algorithm. Three methods were introduced for creating separable clusters of gaze directions to improve the performance of the classification. In our first method we perform a summation of the gaze directions. In our second method we remove the head pose component. In the third method we use an estimate of the camera locations to project the gaze vectors onto a hyperplane that is created by the target's pose location. Our unsupervised approach to gaze aversion classification is one that could be extended. More features, such as temporal information, could be added to improve the classification performance. We also note that the three different methods could be used with different clustering algorithms apart from SC, such as hierarchical clustering methods. The first two methods can also be applied to video files where the camera's intrinsic parameters are unknown. The performance of our approach should improve as appearance-based methods of gaze tracking improve.

Acknowledgment

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

[1] M. Argyle and M. Cook, *Gaze and Mutual Gaze*. Cambridge University Press, 1976.

[2] S. R. Langton, R. J. Watt, and V. Bruce, "Do the eyes have it? Cues to the direction of social attention," *Trends in cognitive sciences*, vol. 4, no. 2, pp. 50–59, 2000.

[3] D. Keltner, "Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame.," *Journal of Personality and Social Psychology*, vol. 68, no. 3, p. 441, 1995.

[4] M. R. Key, *The Relationship of Verbal and Nonverbal Communication*, vol. 25. Walter de Gruyter, 1980.

[5] T. Baltru, P. Robinson, L.-P. Morency, and others, "OpenFace: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, IEEE, 2016.

[6] E. Wood, T. Baltruaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling, "Rendering of Eyes for Eye-Shape Registration and Gaze Estimation," pp. 3756–3764, IEEE, Dec. 2015.

[7] C. Alzate and J. Suykens, "Multiway Spectral Clustering with Out-of-Sample Extensions through Weighted Kernel PCA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 335–347, Feb. 2010.

[8] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta Psychologica*, vol. 26, pp. 22–63, Jan. 1967.

[9] A. M. Glenberg, J. L. Schroeder, and D. A. Robertson, "Averting the gaze disengages the environment and facilitates remembering," *Memory & cognition*, vol. 26, no. 4, pp. 651–658, 1998.

[10] D. C. Richardson and R. Dale, "Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension," *Cognitive science*, vol. 29, no. 6, pp. 1045–1060, 2005.

[11] L.-P. Morency, C. M. Christoudias, and T. Darrell, "Recognizing gaze aversion gestures in embodied conversational discourse," in *Proceedings of the 8th International Conference on Multimodal Interfaces*, pp. 287–294, ACM, 2006.

[12] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4511–4520, 2015.

[13] K. Liang, Y. Chahir, M. Molina, C. Tijus, and F. Jouen, "Appearance-based gaze tracking with spectral clustering and semi-supervised gaussian process regression," in *Proceedings of the 2013 Conference on Eye Tracking South Africa*, pp. 17–23, ACM, 2013.

[14] L.-P. Morency, A. Quattoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference On*, pp. 1–8, IEEE, 2007.

[15] D. Bohus, C. W. Saw, and E. Horvitz, "Directions robot: In-the-wild experiences and lessons learned," in *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, pp. 637–644, International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[16] S. O. Ba and J.-M. Odobez, "Recognizing visual focus of attention from head pose in natural meetings," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 39, no. 1, pp. 16–33, 2009.

[17] S. Sheikhi and J.-M. Odobez, "Recognizing the visual focus of attention for human robot interaction," in *International Workshop on Human Behavior Understanding*, pp. 99–112, Springer, 2012.

[18] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[19] A. Y. Ng, M. I. Jordan, Y. Weiss, and others, "On spectral clustering: Analysis and an algorithm," in *NIPS*, vol. 14, pp. 849–856, 2001.

[20] R. van Son, W. Wesseling, E. Sanders, and H. van den Heuvel, "The IFADV Corpus: A Free Dialog Video Corpus.," in *LREC*, pp. 501–508, 2008.

[21] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[22] R. Wagner, O. Birbach, and U. Frese, "Rapid development of manifold-based graph optimization systems for multi-sensor calibration and SLAM," in *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference On*, pp. 3305–3312, IEEE, 2011.