Efficient autism spectrum disorder prediction with eye movement: A machine learning framework

Conference Paper · September 2015

DOI: 10.1109/ACII.2015.7344638

CITATIONS
6

READS
57

6 authors, including:

Bhiksha Raj
Carnegie Mellon University
304 PUBLICATIONS **6,434** CITATIONS
SEE PROFILE

Li Yi
Peking University
22 PUBLICATIONS **123** CITATIONS
SEE PROFILE

Ming Li
Duke Kunshan University
93 PUBLICATIONS **977** CITATIONS
SEE PROFILE

Efficient Autism Spectrum Disorder Prediction with Eye Movement: A Machine Learning Framework

Wenbo Liu*†‡
*SYSU-CMU Joint Inst. of Eng.
Sun Yat-sen University
Guangzhou, China 510006

Zhiding Yu
†Dept. of Electrical & Computer Eng.
Carnegie Mellon University
Pittsburgh, PA 15213

Bhiksha Raj
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Li Yi
Department of Psychology
Peking University
Beijing, China 100871

Xiaobing Zou
The Third Affiliated Hospital
Sun Yat-sen University
Guangzhou, China 510620

Ming Li*‡
‡SYSU-CMU Joint Research Inst.
Shunde, China 528300
liming46@mail.sysu.edu.cn

***Abstract*—We propose an autism spectrum disorder (ASD) prediction system based on machine learning techniques. Our work features the novel development and application of machine learning methods over traditional ASD evaluation protocols. Specifically, we are interested in discovering the latent patterns that possibly indicate the symptom of ASD underneath the observations of eye movement. A group of subjects (either ASD or non-ASD) are shown with a set of aligned human face images, with eye gaze locations on each image recorded sequentially. An image-level feature is then extracted from the recorded eye gaze locations on each face image. Such feature extraction process is expected to capture discriminative eye movement patterns related to ASD. In this work, we propose a variety of feature extraction methods, seeking to evaluate their prediction performance comprehensively. We further propose an ASD prediction framework in which the prediction model is learned on the labeled features. At testing stage, a test subject is also asked to view the face images with eye gaze locations recorded. The learned model predicts the image-level labels and a threshold is set to determine whether the test subject potentially has ASD or not. Despite the inherent difficulty of ASD prediction, experimental results indicates statistical significance of the predicted results, showing promising perspective of this framework.**

***Keywords*—autism spectrum disorder; eye tracking; bag-of-words; support vector machine**

I. Introduction

Rate of autism spectrum disorder (ASD) has risen sharply in the past several years, reaching 1 in 68 in the United States [1]. Despite the fact that existing assessment methods show high validity, current ASD diagnostic approaches are both time and labour consuming. In particular, the diagnostic instruments have been designed to measure impairments mainly in: (1) language and communication; (2) reciprocal social interactions; and (3) restricted, repetitive behaviors. The most widely used instruments include the Autism Diagnostic Observation Schedule-Generic (ADOS-G) [2] and the revised version ADOS-2 [3]. These approaches require the accompany and administration of clinically trained professionals and the whole process can take up to 90 minutes. The interactive, human-in-loop nature of these tests not only increase unnecessary costs, but also reduces the chance of early diagnosis.

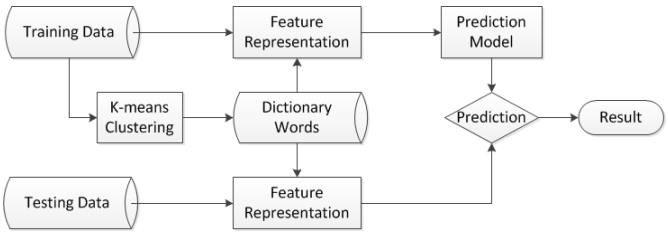


Fig. 1. The proposed machine learning based ASD prediction framework.

Behavioral studies found that ASD individuals have impairments in recognizing human faces, leading to atypical face processing [4]. Although there exists considerable controversy regarding whether ASD individuals also scan faces differently, recent studies indeed indicate evidence of different eye movement patterns from ASD individuals [5], [6]. The way how ASD individuals scan faces has been studied by a number of previous literatures with eye tracking techniques. Existing studies have consistently indicated that children and adults show reduced visual attentions to faces compared to their typically developed (TD) counterparts [7]. While these closely related studies form the fundamental bases of this research, most of them are only restricted to discovering statistical significant ASD patterns and few of them considered prediction tasks. A major contribution of this paper is that we propose a machine learning based framework (see Fig. 1) on face scanning pattern analysis as an alternative ASD measurement. Compared with traditional instruments such as ADOS-G and ADOS-2, the proposed framework requires much less human interaction and expertise. We do not argue that such framework can completely replace traditional ones. Rather, it can be regarded as a supplement that benefits earlier and more accurate ASD diagnosis. There are studies that also use machine learning to optimize the diagnosis process [8], [9], [10]. These studies, however, do not change the highly interactive essence of traditional diagnosis procedure.

Another major contribution includes the improvement over

978-1-4799-9953-8/15/$31.00 ©2015 IEEE

existing approaches such as area of interest (AOI) and iMAP[11]. The AOI approach was widely used by many to conduct analysis on the face scanning patterns. It seeks to measure eye fixations that fall within a predefined area of interest, typically including the AOIs of eyes, nose and mouth. With the defined AOIs, one is able to statistically estimate the frequency counts of eye fixations on different face areas. A common problem with the AOI approach is that it tends to lump fixations to a relatively large area without further discrimination. The boundary of AOI is often determined empirically and is influenced by the semantic meaning of face parsing without statistical and psychological justification, while a subject's visual attention could in fact be largely influenced by certain sub-AOI regions highly responsive to human brains as mid-level visual features. The iMAP approach was proposed as an alternative method which supplements the AOI approach regarding such issues. iMAP uses a Gaussian kernel to spatially smooth each fixation map and operates at pixel level to compare different conditions or groups with statistical normalization. It is, therefore, able to reveal discriminative spatial differences at a much finer spatial resolution.

A key feature extraction step in our work is using k-means to cluster the eye gaze and divide the face into different sub-regions. This step shares certain similarity with both AOI and iMAP, in the sense that it partitions spatial face regions like AOI, but is more data-driven and returns more flexible partitions like iMAP. The advantage of such data-driven strategy is that it generates partitions with statistical importance. Intuitively, the cluster centroids are the most representative face scanning "hot spots" found by k-means. Our proposed work not only targets the eye gaze coordinates on spatial domain, but also exploits eye movement on motion domain. Besides coordinates, we also conduct k-means on the difference of consecutive eye gaze coordinates. The expectation is that the magnitude and the direction of eye motion may also reveal certain ASD evidence. To our best knowledge, very few previous literatures have investigated from this perspective.

Following the k-means clustering is the "bag-of-words" (BoW) histogram feature representations of both coordinates and motion magnitudes. The cluster centroids returned by k-means are referred to as "dictionary words" in the "bag-of-words" model. Such histogram representation is basically an orderless frequency encoding of both the participant's visual attention on each part of a face, and the motion magnitudes/directions. In the experiments, we will comprehensively evaluate the prediction performance of the above features as well as a fusion of these methods.

II. Dataset Configuration

To conduct experiments, we consider the two eye movement datasets from [5] and [6]. For both datasets, the eye gaze of each subject were recorded by a Tobii T60 eye tracker. The eye tracker has sample rate of 60 Hz and a screen resolution of 1024×768 pixels. A set of face images (700×500) are displayed on the screen and eye gaze of each subject is automatically estimated, returning a set of projected coordinates on the screen. While it is possible that some of the coordinates can fall out of the 700×500 image domain, we only consider the majority of coordinates that are within the domain.

The first dataset [5] targets the ASD behavioral analysis on children and contains three groups of Chinese children: 20 ASD children, 21 age-matched typically developing (TD) children, and 20 IQ-matched TD children. The second dataset [6] on the other hand focuses on adolescents and young adults, including 19 ASD, 22 IQ-matched intellectually disabled (ID), and 28 age-matched TD adolescents/young adults. The readers can kindly refer to both literatures for more details on data collection setups[1].

III. The Proposed Framework

Feature representation is an indispensable step for predictions. A feature is an individual measurable property of a phenomenon being observed, while feature representation refers to the numerical (such as scalars, vectors or matrices) or structural (such as graphs or strings) description of such measurable properties. The fundamental principle of feature representation is choosing informative features as well as finding good representations. Feature representation also shows certain connections to past studies on ASD face scanning, as statistical analysis in these studies more or less fall into the domain of seeking informative features.

Our key focus in this paper lies in conducting eye movement analysis at (face) image level. Fig. 2 illustrates the overall dataset infrastructure from subjects to image-level features. In particular, a single feature is extracted from the eye movement data recorded per face image per subject. Each feature is labeled either positive or negative based on the identity (ASD/non-ASD) of the subject. Once features are extracted, a prediction model is trained and is tested at image-level for the new test subject.

A. Bag-of-Words Feature on Eye Gaze Coordinates

The first feature we consider is the BoW histogram representation on the gaze coordinates. The BoW model originally came from the linguistic community [12] and has ever since been a very popular feature representation framework with wide applications in Natural Language Processing, information retrieval [13] and computer vision [14]. The reason why such model is called "bag-of-words" is because a sentence or a document can be represented as the bag (multi-set) of its words, disregarding grammar and even word order but keeping multiplicity.

Similar analogies can be made here as we treat the centers of concentrated visual attentions as dictionary words, while the sequence of coordinates per image per subject as one document. An atypical frequency distribution of gaze on different parts of a face image can be a strong evidence of reduced visual attention. To discover important spatial regions for eye movement patterns, we use the k-means algorithm to cluster the recorded eye gaze coordinates from all participants

[1] In the following paper, we will denotes these two datasets respectively as "Child" and "Adult" for short.

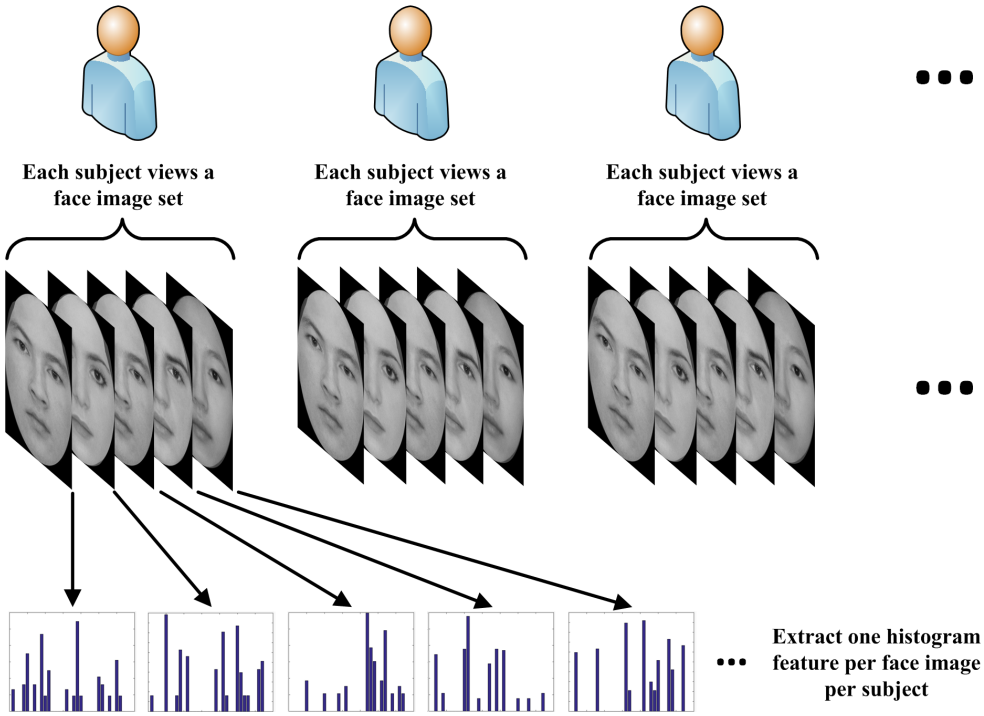


Fig. 2. The dataset infrastructure.

in the training set, and divide the face image into different sub-regions. The resulting output is a set of cell-like spatially partitioned face regions indicating clusters of gazes that are relatively more concentrated. Four partitioning examples with dictionary numbers respectively equal to 16, 32, 48 and 64 are shown in Fig. 3.

We consider two types of histogram representations:

Hard histogram: The generated histogram feature is a sparse vector of occurrence counts of every cluster. That is, a histogram indicating the frequencies of eye gaze on different parts of a face. Upon obtaining the frequency counts, a histogram is normalized such that the sum of elements is equal to 1.

Soft histogram: The histogram feature is an accumulation of soft memberships to different dictionary words. A membership is a value between $(0, 1)$ measuring the belongingness to certain dictionary. Let $\mathbf{x}_{i,j,n}$ denote the nth eye gaze coordinate of the ith subject on j image, The membership to the kth dictionary is computed as:

$$u^k_{i,j,n} = \frac{1/\|\mathbf{x}_{i,j,n} - \mathbf{d}_k\|_2^2}{\sum_{k=1}^{K} 1/\|\mathbf{x}_{i,j,n} - \mathbf{d}_k\|_2^2} \tag{1}$$

It is very easy to verify that $\sum_{k=1}^{K} u^k_{i,j,n} = 1$. Let $\mathbf{u}_{i,j,n} = [u^1_{i,j,n}, ..., u^K_{i,j,n}]$, the soft histogram can be computed as:

$$\mathbf{h}_{i,j} = \frac{1}{N_j} \sum_{n=1}^{N_j} \mathbf{u}_{i,j,n}. \tag{2}$$

The soft histogram returns a softer frequency counting than hard histogram since it considers all dictionaries instead of hard assigning to the closest one. Using a soft histogram may benefit the counting of eye gazes that fall right on the border of two regions.

We also consider a simple yet very useful technique called the **square root representation**. The square root regularization simply takes the square root of each entry in the histogram:

$$\begin{aligned} \sqrt{\mathbf{h}_{i,j}} &:= [\sqrt{h^1_{i,j}}, ..., \sqrt{h^K_{i,j}}] \\ s.t. \quad & \sum_{k=1}^{K} (\sqrt{h^k_{i,j}})^2 = 1 \end{aligned} \tag{3}$$

This projects every histogram onto the unit K-dimensional hypersphere. Such square root representation implicitly leads to the Bhattacharyya kernel when we consider the inner product of two transformed histograms:

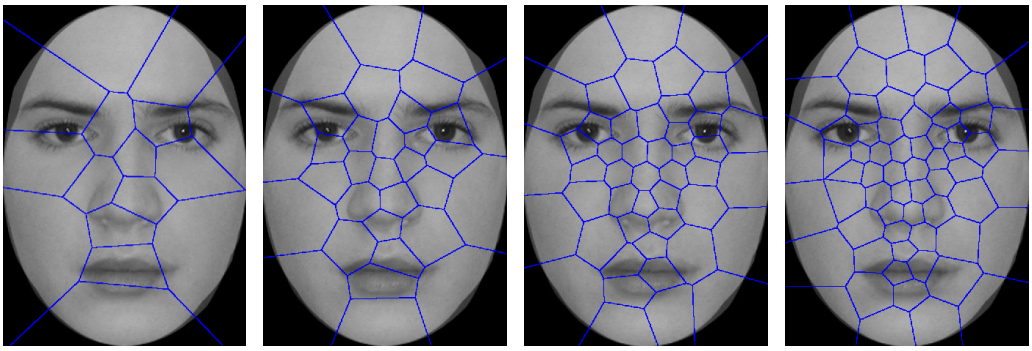
$$K_S(\mathbf{h}_i, \mathbf{h}_j) = \sum_{k=1}^{K} \sqrt{\mathbf{h}^k_i \mathbf{h}^k_j}. \tag{4}$$

The square root representation is very effective in suppressing noise and boosting classification performance. In the experiment we include such transform as feature preprocessing.

B. Bag-of-Word Feature on Eye Motion

The second feature we consider is the BoW representation on eye gaze motion. The motion vector of eye gaze is computed as:

$$\mathbf{m}_{i,j,n} = \mathbf{x}_{i,j,n+1} - \mathbf{x}_{i,j,n} \tag{5}$$



(a) (b) (c) (d)

Fig. 3. An illustration of partitioned face regions by k-means with different cluster (dictionary) numbers. (a) $K = 16$. (b) $K = 32$. (c) $K = 48$. (d) $K = 64$. The illustrated face is an averaged face in order to protect the privacy of test subject.

After getting the motion vectors, the BoW histogram features are extracted in the same as described in the Section III-A.

C. Image-Level Prediction

We label all the obtained image-level BoW histogram features by the identity (ASD or non-ASD) of subjects. A binary classier is then trained on these labeled features. Support vector machine (SVM) is a widely used classifier for its excellent performance. It learns a linear decision boundary such that the margin separating the positive data and negative data is maximized. In our work, we adopt SVM to learn the prediction model for the BoW features on both eye gaze coordinates and motions.

Linear SVM only learns a linear decision boundary. The obtained histogram features, however, are often not linearly separable due to the complexity of face scanning behaviors. We consider kernel SVM to introduce nonlinearity to the decision boundary. There are many types of kernel functions. In this study we choose the popular radial basis function (RBF) kernel for its good performance.

D. Subject-Level Prediction

Image-level predictions are less robust due to the limitation of information conveyed by every image-wise test. A subject level prediction on the other hand is what one ultimately desires. Therefore, we ensemble image-level predictions to finalize subject-level predictions by fixing a global threshold. We consider the following two ways of ensemble strategies:

Soft prediction: The RBF kernel SVM gives each image-level testing feature a soft prediction score. Suppose each feature is re-denoted as $\mathbf{h}_n$ with a single, global index. Also let $y_n \in \{-1, 1\}$ denotes the label of feature $\mathbf{h}_n$, $\mathbf{w}$ and b denote the learned parameters defining the decision hyperplane. The soft prediction score is computed as:

$$\begin{aligned} soft\_score(n) &= \mathbf{w}^\top \Phi(\mathbf{h}_n) + b \\ &= \sum_m \alpha_m y_m K(\mathbf{h}_m, \mathbf{h}_n) + b. \end{aligned} \tag{6}$$

The top row of (6) gives an intuitive geometric interpretation of the score. It is basically the functional margin of a kernelized feature and the decision boundary. In practice however, the prediction score is obtained by solving the dual problem of SVM, resulting in the second row where α_i are the introduced lagrange multipliers and $K(\cdot, \cdot)$ is the kernel function. The subject-level mean score is defined as:

$$subject\_score(i) = \frac{1}{|S_i|} \sum_{n \in S_i} soft\_score(n), \tag{7}$$

where S_i corresponds to the set of global indexes belonging to the ith subject.

Hard prediction: The RBF kernel SVM gives each image-level testing feature with a $\{0, 1\}$ hard score:

$$hard\_score(n) = \begin{cases} 1, & \text{if} \quad soft\_score(n) > 0 \\ 0, & \text{else} \end{cases} \tag{8}$$

Again the subject-level mean score is defined as:

$$subject\_score(i) = \frac{1}{|S_i|} \sum_{n \in S_i} hard\_score(n), \tag{9}$$

The subject level prediction for both methods is determined with a global threshold T:

$$subject\_score(i) \underset{non-ASD}{\overset{ASD}{\gtrless}} T \tag{10}$$

IV. Experiment Configuration

We adopt the leave-one-out strategy to choose one subject as testing subject each time and leave the rest of the participants as training subjects. By doing this we divide the obtained image-level BoW features into two portions: a training set and a test set. Such leave-one-out strategy is repeated for every participant. For all experiments, the parameters of SVM, which are γ and C, were empirically optimized and fixed for every ROC curve.

A. Baseline-1: N-Gram Model

Besides the BoW representation of eye gaze coordinates, we also use the N-Gram modeling to consider temporal correlation between subsequent gaze coordinates. Given a set of dictionary words trained by k-means, each coordinate is assigned to

one of the dictionary words. Therefore, the whole set of coordinates is transformed in to a set of discrete cluster labels indicating the dictionary membership of each coordinate. Instead of having a single feature per image per subject, we currently have a label sequence per image per subject. We use SRILM [15] which respectively takes the positive training sequences, the negative training sequences and the testing sequences (of the one test subject) as three groups, and outputs two log likelihood scores $l\_pos(i)$ and $l\_neg(i)$ which respectively indicates how likely the test subject (the ith subject) is belonging to the positive training group and the negative training group. The final subject-level score is normalized as:

$$subject\_score(i) = \frac{|l\_pos(i) - l\_neg(i)|}{|l\_pos(i)| + |l\_pos(i)|} \qquad (11)$$

B. Baseline-2: BoW Feature from AOI Dictionaries

AOIs can be regarded as alternative dictionaries with semantic meanings. In our experiment, AOIs are defined separately for each face image, where we segment the face into several semantic parts: face, nose, mouth, left eye and right eye. Like dictionary words extracted by k-means, a BoW histogram can also be extracted from AOI dictionaries by computing and normalizing the frequency counts of eye gaze coordinates falling onto each semantic part. While previous literatures only conduct statistical significance analysis based on AOI frequency counts and are not directly comparable, we choose to plug it into our prediction framework as a baseline feature.

C. Evaluation Benchmarks

In the experiment, we use the following benchmarks to quantitatively evaluate the prediction performance:

ROC curve: We vary the global threshold and calculate the corresponding set of (subject-level) true positive rates versus false positive rates.

Area under curve (AUC): The total area under the ROC curve versus the whole area.

Accuracy: The number of correctly predicted subjects versus the total number of subjects

V. Experimental Results

In this section, we report the comprehensive evaluations of our proposed features and methods on the two datasets.

A. BoW Feature on Eye Gaze Coordinates

We first evaluate the eye coordinate BoW feature. The image-level prediction is chosen to be soft prediction. Figure 4 shows results of the eye gaze BoW representation of eye gaze coordinates on the Adult dataset. We vary the number of dictionary numbers K and test both the hard and soft histograms to comprehensively evaluate our proposed features. In addition, we measure the AUC versus the number of dictionaries, and select the two best ROC curves respectively from results of hard histogram and results of soft histogram, measuring the corresponding accuracies. One could observe that the soft histogram in general performs better than hard

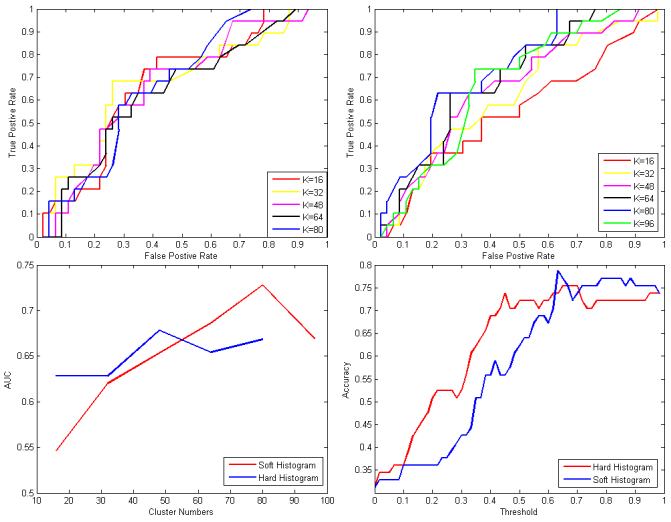


Fig. 4. Results of hard BoW feature and soft BoW feature on Eye gaze coordinates on the Adult dataset. The top left image corresponds to hard histograms, while the top right image corresponds to soft histograms. The bottom left image are the AUC values from both hard histograms and soft histograms. Finally, two best ROC curves are selected with their accuracies measured.

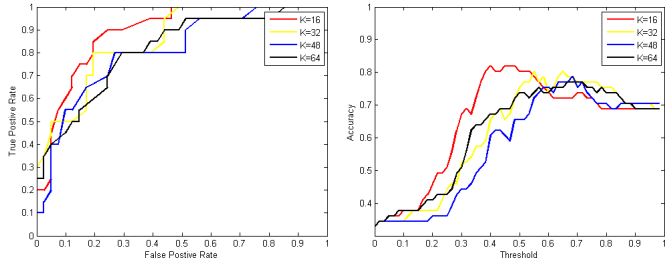


Fig. 5. Results of soft BoW feature on Eye gaze coordinates on the Child dataset. Left image: ROC curves with different dictionary numbers. Right image: Corresponding accuracies with different dictionary numbers.

histograms, showing the benefit brought by soft counting. In the following experiment, we will fix all the extracted histograms as soft ones.

We also demonstrate the prediction results on Child dataset using the soft histogram and soft prediction. Figure 5 shows the corresponding ROC curves and AUC values. One could see that the experiment shows very promising results, indicating strong evidence of statistical significance of certain discriminative eye movement patterns captured by the proposed feature extraction method.

B. Soft Prediction vs. Hard Prediction

We respectively select the optimal dictionary numbers for Adult dataset and Child dataset, from soft BoW results reported in Section V-A. We then use soft BoW representation of eye gaze coordinates, and conduct tests on the two datasets with both soft and hard predictions. Fig. 6 shows the results and one could see that soft prediction does boost the prediction performance. In subsequent experiments, we will fix the prediction method as soft prediction.

978-1-4799-9953-8/15/$31.00 ©2015 IEEE

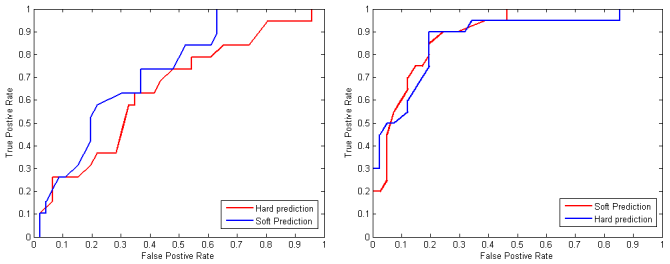


Fig. 6. Results of soft BoW from eye gaze coordinates, with both prediction methods. Left image: ROC curves of soft and hard prediction on Adult dataset. Right image: ROC curves of soft and hard prediction on Child dataset.

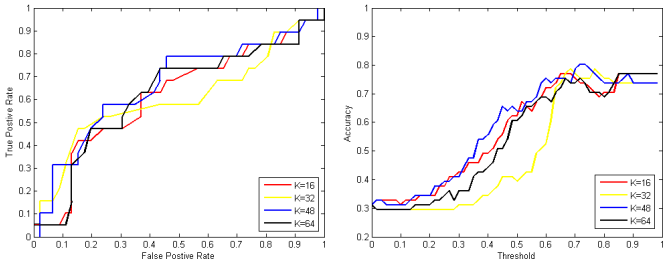


Fig. 7. Results of soft BoW from Eye gaze motions on the Adult dataset. Left image: ROC curves with different dictionary numbers. Right image: Corresponding accuracies with different dictionary numbers.

C. Bow Feature on Eye Gaze Motions

We also test soft BoW from eye gaze motions on Adult and Child datasets. with results shown in Fig. 7 and Fig. 8. Results indicate that motion feature is also very discriminative.

D. Feature Fusion and Comparison with N-Gram and AOI

We finally consider feature-level fusion of both Eye gaze coordinates motions. The ROC and accuracy curves of both BoW features as well as fused one are shown in Fig. 9 and 10. We also show results of two baselines: N-Gram and BoW with AOI. The best accuracy and AUC of all comparing methods are listed in Table I. The table indicates several aspects: 1. The fused (concatenated) feature can further boost the performance since two features are complementary to each other. 2. The proposed method significantly outperforms baseline methods.

We conducted statistical significance test on subject-level soft prediction scores, measuring p-values. The test results are normalized and therefore eliminate the chance influence from unbalanced test data. The p-value is around 0.5 on Adult and lower than 0.001 on Child. The latter particularly shows strong statistical significance. We believe Child dataset to some extent is more important since early diagnosis of ASD maximizes the gain of early intervention.

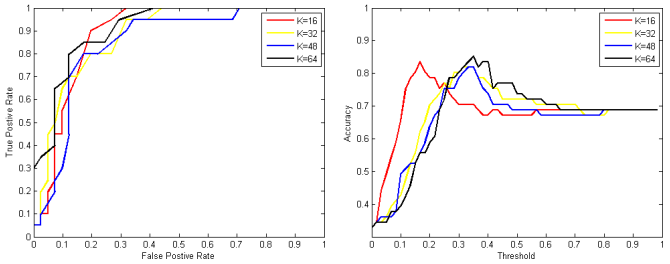


Fig. 8. Resutls of soft BoW from Eye gaze motions on the Child dataset. Left image: ROC curves with different dictionary numbers. Right image: Corresponding accuracies with different dictionary numbers..

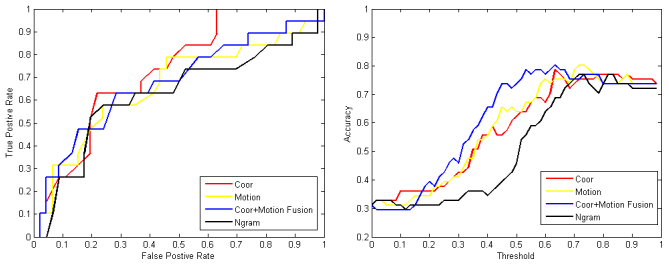


Fig. 9. Results of different proposed methods on the Adult dataset. Left image: ROC curves with different methods. Right image: Corresponding accuracies with different methods.

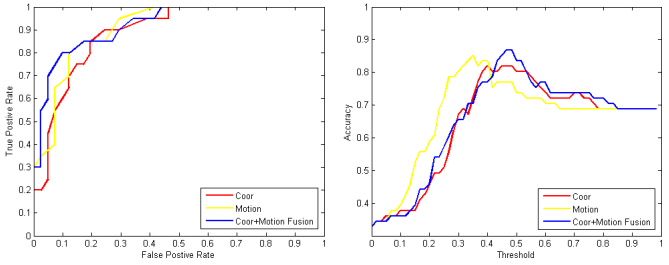


Fig. 10. Results of different proposed methods on the Child dataset. Left image: ROC curves with different methods. Right image: Corresponding accuracies with different methods.

VI. Conclusion

In this paper, we proposed a machine learning framework for ASD prediction based on face scanning eye movement data. We also proposed a comprehensive set of effective feature extraction methods, prediction frameworks, as well as corresponding scoring frameworks. Despite the great challenge of this problem, we have achieved promising results on two ASD datasets, particularly on the child set. Experimental results indicate the effectiveness and potential future value of the proposed methods.

VII. Acknowledgement

This research is partially funded by the National Natural Science Foundation of China (61401524), Natural Science Foundation of Guangdong Province (2014A030313123), CMU-SYSU Collaborative Innovation research Center Foundation (35321.2.8.1011475), and SYSU-CMU Shunde International Joint Research Institute.

TABLE I
Quantitative prediction results of different methods

		Coor	Motion	Fusion	AOI	N-Gram
Adult	**AUC**	**0.7277**	0.6636	0.6773	-	0.6483
	Accuracy	0.7869	**0.8033**	**0.8033**	-	0.7705
Child	**AUC**	0.8902	0.9061	**0.9207**	0.8208	0.5561
	Accuracy	0.8197	0.8525	**0.8689**	0.7868	0.7213

REFERENCES

[1] M. Wingate, R. S. Kirby, S. Pettygrove, C. Cunniff, E. Schulz, T. Ghosh, C. Robinson, L.-C. Lee, R. Landa, J. Constantino *et al.*, "Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, united states, 2010," *MMWR Surveillance Summaries*, vol. 63, no. 2, 2014.

[2] C. Lord, S. Risi, L. Lambrecht, E. H. Cook Jr, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, "The autism diagnostic observation schedulełgeneric: A standard measure of social and communication deficits associated with the spectrum of autism," *Journal of autism and developmental disorders*, vol. 30, no. 3, pp. 205–223, 2000.

[3] K. Gotham, S. Risi, A. Pickles, and C. Lord, "The autism diagnostic observation schedule: revised algorithms for improved diagnostic validity," *Journal of autism and developmental disorders*, vol. 37, no. 4, pp. 613–627, 2007.

[4] S. Weigelt, K. Koldewyn, and N. Kanwisher, "Face identity recognition in autism spectrum disorders: a review of behavioral studies," *Neuroscience & Biobehavioral Reviews*, vol. 36, no. 3, pp. 1060–1084, 2012.

[5] L. Yi, Y. Fan, P. C. Quinn, C. Feng, D. Huang, J. Li, G. Mao, and K. Lee, "Abnormality in face scanning by children with autism spectrum disorder is limited to the eye region: Evidence from multi-method analyses of eye tracking data," *Journal of vision*, vol. 13, no. 10, p. 5, 2013.

[6] L. Yi, C. Feng, P. C. Quinn, H. Ding, J. Li, Y. Liu, and K. Lee, "Do individuals with and without autism spectrum disorder scan faces differently? a new multi-method look at an existing controversy," *Autism Research*, vol. 7, no. 1, pp. 72–83, 2014.

[7] T. Falck-Ytter and C. von Hofsten, "How special is social looking in asd: a review." *Progress in brain research*, no. 189, pp. 209–22, 2011.

[8] D. Bone, M. S. Goodwin, M. P. Black, C.-C. Lee, K. Audhkhasi, and S. Narayanan, "Applying machine learning to facilitate autism diagnostics: Pitfalls and promises," *Journal of autism and developmental disorders*, pp. 1–16, 2014.

[9] J. Kosmicki, V. Sochat, M. Duda, and D. Wall, "Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning," *Translational psychiatry*, vol. 5, no. 2, p. e514, 2015.

[10] M. Duda, J. Kosmicki, and D. Wall, "Testing the accuracy of an observation-based classifier for rapid detection of autism risk," *Translational psychiatry*, vol. 4, no. 8, p. e424, 2014.

[11] R. Caldara and S. Miellet, "imap: a novel method for statistical fixation mapping of eye movement data," *Behavior research methods*, vol. 43, no. 3, pp. 864–878, 2011.

[12] Z. S. Harris, "Distributional structure." *Word*, 1954.

[13] J. Sivic and A. Zisserman, "Efficient visual search of videos cast as text retrieval," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 4, pp. 591–606, 2009.

[14] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2. IEEE, 2005, pp. 524–531.

[15] A. Stolcke *et al.*, "Srilm-an extensible language modeling toolkit." in *INTERSPEECH*, 2002.

978-1-4799-9953-8/15/$31.00 ©2015 IEEE