This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2016.2543681, IEEE Journal of Selected Topics in Signal Processing

Social Interaction Assistant: A Person-Centered Approach to Enrich Social Interactions for Individuals with Visual Impairments

Sethuraman Panchanathan, *Fellow, IEEE,* Shayok Chakraborty, *Member, IEEE* and Troy McDaniel, *Member, IEEE*

***Abstract*—Social interaction is a central component of human experience. The ability to interact with others and communicate effectively within an interactive context is a fundamental necessity for professional success as well as personal fulfillment. Individuals with visual impairment face significant challenges in social communication, which if unmitigated, may lead to lifelong needs for extensive social and economic support. Unfortunately, today's multimedia technologies largely cater to the needs of the "able" population, resulting in solutions that mostly meet the needs of that community. Individuals with disabilities (such as visual impairment) have largely been absent in the design process, and have to adapt themselves (often unsuccessfully) to available solutions. In this paper, we propose a social interaction assistant for individuals who are blind or visually impaired, incorporating novel contributions in: (i) person recognition through batch mode active learning, (ii) reliable multi-modal person recognition through the conformal predictions framework and (iii) facial expression recognition through topic models. Moreover, individuals with visual impairments often have specific requirements that necessitate a personalized, adaptive approach to multimedia computing. To address this challenge, our proposed solutions place emphasis on understanding the individual user's needs, expectations and adaptations towards designing, developing and deploying effective multimedia solutions. Our empirical results demonstrate the significant potential in using person centered multimedia solutions to enrich the lives of individuals with disabilities.**

***Index Terms*—Social Interaction Assistant, Machine Learning, Computer Vision, Person-centered Multimedia Computing**

I. Introduction

SOCIAL interaction refers to any form of mutual communication between two individuals (dyadic interactions) or between an individual and a group (group interactions). A strong set of social skills is important for a successful and productive life. For example, they help us make new friends, or make good first impressions at job interviews. Sociologists believe that social interactions are the underpinnings of our modern society, and are essential for social development and acceptance of an individual within our society. Such communications typically involve many types of sensory and motor activities, as deemed necessary by the participants of the interaction. Social, behavioral and developmental sociologists emphasize that the ability of individuals to effectively employ expressive behavior is essential for the social and interpersonal functioning of society. Human interactions are a combination of verbal and non-verbal cues, where the latter implies implicit communication cues that use prosody, body kinesis and facial movements to convey information. In everyday interactions, people communicate so effortlessly through verbal and non-verbal cues that they are not aware of the complex interplay of their voice, face and body in establishing a smooth communication channel. While spoken language plays an important role in communication, speech accounts for only 35% of the interpersonal exchanges. Nearly 65% of all information communication happens through non-verbal cues [1]. As depicted in Figure 1, approximately 48% of the communication is through visual encoding of face, body kinesis and posture, while the rest is encoded in the prosody (intonation, pitch, pace and loudness of voice) [2].

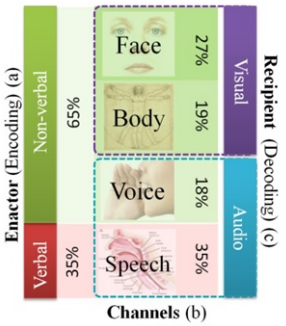


Fig. 1. Relative importance of a) verbal vs. non-verbal cues, b) four channels of non-verbal cues and c) visual vs. audio encoding and decoding of bilateral human interpersonal communicative cues. Based on the meta-analysis in [3].

Since most of the non-verbal cues (eye-gaze, head nod, body mannerisms, facial expressions) are perceived visually, people with visual impairments get deprived of these vital communicative cues and face significant challenge in interacting with their sighted peers. Limited access to non-verbal cues can create mis-communications, leading to embarrassing social situations, increasing the likelihood of social isolation.

According to current statistics, approximately 10 percent of the world's population or roughly 650 million people, live with some form of disability. In the U.S., 36 million people have at least one disability, which is about 12 percent of the total U.S. population. Since January 2011, this number has rapidly grown considering that 10,000 baby boomers turn 65 every day, each of whom will encounter sensory, physical

The authors are with the Center for Cognitive Ubiquitous Computing (CUbiC) at Arizona State University, Tempe, AZ 85281
E-mail: {panch, shayok.chakraborty, troy.mcdaniel}@asu.edu

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2016.2543681, IEEE Journal of Selected Topics in Signal Processing

and/or cognitive impairments during their activities of daily living[1]. Despite these alarming figures, today's multimedia technology are largely geared toward the "able" population, with little consideration to accessibility. Accessibility within commercial products are often an after-thought, rather than being an integral part of the design from inception. Special-purpose assistive technology and software are available, but they come with a hefty price tag due to smaller market segments and potential additional regulatory costs. Thus, there is a pronounced need for the development of multimedia technologies for the large percentage of the disabled population. However, this necessitates a thorough and critical analysis, as disabilities are diverse and accommodating every user via a universal design may render overly complicated systems deemed unusable by most [4].

To address this fundamental challenge, the Center for Cognitive Ubiquitous Computing (CUbiC) at Arizona State University[2] has dedicated its efforts toward the development of multimedia computing technologies for the disabled population [5]. An exemplar project is the design and development of a social interaction assistant to enrich the interaction experience of individuals with visual impairments. We adopt a person-centered approach in our solutions that caters to the needs, preferences and behavior of individual users toward the design of multimedia systems. Such an approach is of paramount importance when designing accessible technologies given the large diversity of disabilities and the variation encountered within specific disabilities. An important philosophy of person-centered multimedia computing (PCMC) is co-adaptation [6][7]; it is the bidirectional interaction in which both the user and the system learn and adapt together over time through continual use. Through co-adaptation, individualized designs may be achieved while maintaining applicability to a broad range of users. Further, since the human and the machine work closely together, co-adaptation can help solve complex challenges.

II. The Social Interaction Assistant (SIA)

A. System Description

The goal of CUbiC's Social Interaction Assistant (SIA) is to enhance the accessibility of social non-verbal cues for individuals who are blind or visually impaired. The system is depicted in Figure 2. It consists of a pair of glasses with a camera discreetly embedded in the nose bridge, as shown in the figure. The incoming video stream is analyzed using machine learning and computer vision algorithms to extract relevant non-verbal cues. This information is then delivered to the user. For the delivery phase, it is important to not overload users with information, particularly through a channel that may be obtrusive to ongoing social interaction, such as audio. To overcome this, information delivery occurs mostly through the sense of touch. The haptic belt (as shown in Figure 2) consists of an array of vibration motors and non-verbal cues are delivered through different vibration patterns. For instance, the location of the vibro-tactile stimulation around the waist indicates the position of the interaction partner and the intensity of the vibration denotes his distance.

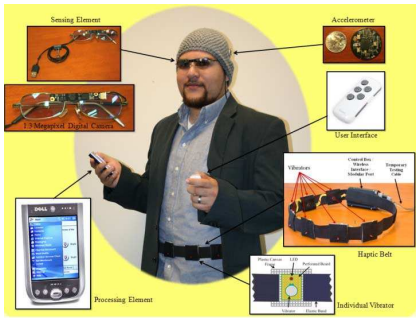


Fig. 2. CUbiC's Social Interaction Assistant.

B. Person-Centeredness in the SIA

To design and develop technologies that enhance social interactions of people who are blind, it was first necessary to gather user input. Two focus groups were conducted, consisting of disability specialists, people who are visually impaired, their parents and family. During focus group sessions, participants freely discussed their needs and any problems encountered in daily life. This was done to ensure person-centeredness in our solutions. From these discussions, a set of eight requirements were identified and a web-based survey was conducted to prioritize needs [8]. Twenty seven people participated in the online survey, of which 16 were blind, 9 had low vision and 2 were sighted specialists in visual impairments. Non-verbal cues were identified that were perceived as being the most needed and important. The results are depicted in Figure 3. We note that non-verbal cues like person recognition (identity) and facial expression recognition were ranked highly among others.

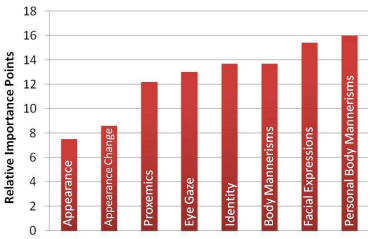


Fig. 3. Results from an online survey. Person recognition and facial expression recognition were ranked high among others.

Both these are well researched problems with significant signal processing challenges. One important challenge is the processing of non-frontal images, i.e. recognizing subjects and their facial expressions from profile images. In our SIA design, a face detection algorithm [9] is used to locate the position of an interaction partner. This information is used to generate vibration signals around the waist, where the location of the vibration indicates the position of the interaction partner. Upon feeling a vibration, the blind user turns his head in the respective direction to center the vibration at his midline. Our

[1]http://www.disabilitystatistics.org/
[2]http://www.cubic.asu.edu

initial pilot testing revealed this feedback loop to be intuitive; in fact, one user described her responses to the vibrations to be an "instant reflex" given the naturalness of the design. Once the user turns in the direction of the vibration, the person recognition algorithm is now faced with the comparatively easier task of recognizing frontal face images. This is a classic example of person-centered multimedia computing where the user and the technology work closely together. The user adapts to the incoming stimuli while the system adapts to the user by utilizing better images for analysis needed for other tasks. Thus, difficult computational challenges, such as fundamental computer vision problems, can be made easier through co-adaptation. This is the fundamental concept of person-centeredness; co-adaptation should occur seamlessly and implicitly to meet the needs and expectations of users.

Further, it is important to note that there are different strata of individuals within the entire population of the visually impaired. Hence it is imperative to design assistive technologies to accommodate the adaptation of features based on an individual user's specific abilities, needs and preferences. There are 4 levels of visual function, according to the International Classification of Diseases[3]: normal vision, moderate visual impairment, severe visual impairment and blindness. Moderate and severe visual impairments are usually grouped under the term "low vision". Blindness can therefore be conceptualized as a spectrum, with each individual residing at a unique location on the scale. Consequently, their needs and expectations from an assistive technology will be different and are expected to change over time. Moreover, some people are blind from birth (congenitally-blind) while some become blind later in life (late-blind) and studies have shown that they vary considerably in terms of their cognitive capabilities [10].

As evident from the users' feedback, the key requirements in the SIA are robust person and facial expression recognition, which pose three fundamental challenges: (i) identifying the salient and exemplar samples from large amounts of unlabeled data in order to train a classification model with much reduced human annotation effort; (ii) computing a measure of confidence or certainty associated with every prediction made by the system and (iii) identifying facial descriptors that can capture a richer space of emotions for improved facial expression recognition. We now elaborate our contributions addressing each of these challenges in the context of the SIA. Further, from the above discussions, it is clear that the SIA technology should be personalized and adaptive, and a generalized design will not cater to the needs of individuals with visual impairments. Our primary contribution in this work is the development of solutions in the context of the SIA, which are adaptive and person-centered; through our experiments, we demonstrate how co-adaptation is an innate component in our solutions.

III. Batch Mode Active Learning (BMAL) for Person Recognition

The camera on the pair of glasses, as depicted in Figure 2, has a high frame rate (25-30 fps) and thus a large number of images will be captured within a very short time span. These face images need to be labeled offline by a human expert to train the underlying classification models, so as to recognize the same subjects accurately in the future. Manual annotation of such a large amount of data is an expensive process in terms of time, labor and human expertise. This necessitates the usage of batch mode active learning (BMAL) frameworks to address this challenge. Such algorithms automatically identify the salient and exemplar instances from large amounts of unlabeled data and tremendously reduce human annotation effort in training classification/regression models.

A. Batch Mode Active Leaning: Background and Rationale

In batch mode active learning, the learner is exposed to a pool of unlabeled instances and it iteratively selects a batch of samples for manual annotation. Brinker [11] proposed a diversity based BMAL technique based on SVMs which queried a diverse batch of samples for annotation. Hoi *et al.* [12] used Fisher information as a measure of model uncertainty and proposed to query a batch of sample which maximally reduced the Fisher information. Guo and Schuurmans proposed an optimization based framework for batch selection by maximizing the log likelihood of the selected instances with respect to the already selected data and minimizing the uncertainty of the unselected unlabeled instances [13]. In our previous work, we proposed an adaptive BMAL algorithm which automatically computed the batch size based on the complexity of the unlabeled data in question [14]. We also developed a BMAL framework based on the convex relaxation of an NP-hard integer quadratic programming problem, with guaranteed bounds on the solution quality [15].

1) Proposed Framework: Consider a scenario where a video stream has been captured by a user who is blind and a BMAL algorithm needs to be applied to select a batch B of salient and exemplar images for manual annotation. The batch size k (number of samples to be selected) is assumed to be known in advance (based on available resources). We now formulate an objective function to address this problem. The set of unlabeled samples is denoted by U_t. Let w^t denote the classification model trained on the available labeled examples L_t. In order to have good generalization performance, the model w^{t+1} trained on $L_t \cup B$ should have high classification confidence on the set of unselected images $U_t - B$. Entropy is taken as a metric to quantify uncertainty, with high entropy values signifying higher uncertainty. We introduce a term in the objective function which ensures that the entropy of the updated model (model trained on the initial training set together with the newly selected batch) on the remaining unselected images is minimal. However, this may result in selection of images only from the high-density regions of the unlabeled set. This is because, from a data geometry point of view, the set of images that are not selected may be dominated by images from high-density regions, constituting a large portion of the data. To address this issue, we append a term which selects images specifically from low-density regions in the data space, i.e. images that have a high measure

[3]http://www.who.int/mediacentre/factsheets/fs282/en/

of distance from the remaining set. More formally, the entropy of the model w^{t+1} on an image x_j is computed as:

$$S(y|x_j, w^{t+1}) = -\sum_{y \in C} P(y|x_j, w^{t+1}) \log P(y|x_j, w^{t+1}) \tag{1}$$

where C is the total number of classes and y is a class label. Also, let ρ_j denote the average distance of an unlabeled image x_j from other images in the unlabeled video U_t. Greater values of ρ_j can be assumed to denote that the image is located in a low-density region. In order to ensure that the objective function is differentiable, we use the Euclidean distance in this work (since Euclidean (L_2) norm is differentiable). Any other differentiable distance metric can also be used. The two conditions described previously can thus be satisfied by defining a performance score function $f(B)$ in the following manner:

$$f(B) = \sum_{i \in B} \rho_i - \lambda \sum_{j \in U_t - B} S(y|x_j, w^{t+1}) \tag{2}$$

where $\lambda \in [0, 1]$ is a trade-off parameter controlling the relative importance of the two terms. The problem therefore reduces to selecting a batch B of k unlabeled images which produces the maximum score f(B). Since the search space is exponentially large, exhaustive search methods are not feasible. We therefore use numerical optimization techniques to solve this problem. Specifically, we define a binary vector M of size equal to the number of elements in the unlabeled pool U_t, $M \in \Re^{|U_t| \times 1}$. Each entry denotes whether the corresponding unlabeled sample will be selected for annotation $(M_i = 1)$ or not $(M_i = 0)$. The batch selection problem in Equation (2) therefore reduces to:

$$\max_{M} \sum_{j \in U_t} \rho_j M_j - \lambda \sum_{j \in U_t} (1 - M_j) S(y|x_j, w^{t+1}) \tag{3}$$

s.t.:

$$M_i \in \{0, 1\}, \forall i \quad \text{and} \quad \sum_{i=1}^{|U_t|} M_i = k \tag{4}$$

The above optimization is an integer programming problem and is NP hard. We therefore relax the constraints to make it a continuous optimization problem:

$$\max_{M} \sum_{j \in U_t} \rho_j M_j - \lambda \sum_{j \in U_t} (1 - M_j) S(y|x_j, w^{t+1}) \tag{5}$$

s.t.:

$$0 \leq M_i \leq 1, \forall i \quad \text{and} \quad \sum_{i=1}^{|U_t|} M_i = k \tag{6}$$

We solve M from this formulation and set the top largest k entries as 1 to approximate the integer solution.

2) Solving the Optimization Problem: The objective function is written in terms of M as:

$$f(M) = \sum_{j \in U_t} \rho_j M_j - \lambda \sum_{j \in U_t} (1 - M_j) S(y|x_j, w^{t+1}) \tag{7}$$

To solve the optimization problem, we use the Quasi Newton method, which assumes that the function can be approximated as a quadratic in the neighborhood of the optimum point and iteratively updates the variable M to guide the functional value towards this local optima. The first derivative of the function and the Hessian matrix of second derivatives need to be computed as parts of the solution procedure. Assuming w^{t+1} remains constant with small iterative updates of M, the first order derivative vector is obtained by taking the partial of the objective with respect to M:

$$\nabla f(M_j) = \rho_j + \lambda S(y|x_j, w^{t+1}) \tag{8}$$

The Hessian starts as an identity matrix and is updated according to the BFGS method [16]. In each iteration, a quadratic programming problem is solved which yields an update direction for M. The step size is obtained using a backtrack line search method based on the Armijo Goldstein equation [16] and guarantee monotonic convergence of the function to the local optimum. The iterations are continued until the change in the value of the objective function is negligible. The final value of M is used to govern the specific points to be selected for the given data stream (by greedily setting the top k entries in M as 1 to recover the integer solution. For further details about the Quasi Newton method, please refer [16]. We note that the objective function is defined in terms of the future classifier w^{t+1}, which is unknown. In the Quasi Newton iterations, w^{t+1} is approximated as the classifier trained on the current training set L_t together with the set of unlabeled points selected in the current iteration (through M), where the label of each selected unlabeled point is assumed to be the same as that of the closest training point in L_t. The pseudocode is outlined in Algorithm 1.

Algorithm 1 Proposed Batch Mode Active Learning Algorithm

Require: Training set L_t, Unlabeled set U_t, parameters λ and batch size k, an initial random guess for M, a stopping threshold α

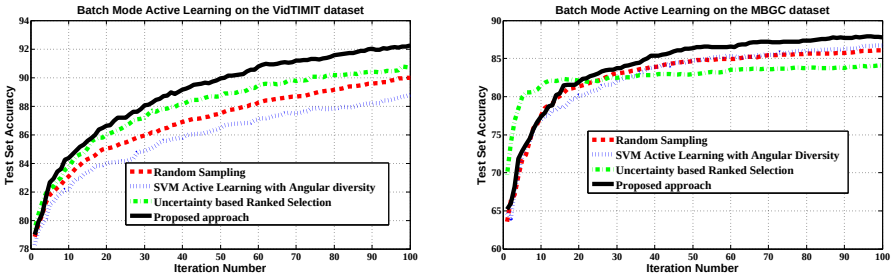
1: Initialize the Hessian matrix H as the identity matrix I
2: Evaluate the objective function $f(M)$ (Equation 7) and the derivative vector $\nabla f(M)$ (Equation 8)
3: **repeat**
4: Solve the QP problem as required by Quasi-Newton: QP(H,$\nabla f(M)$,M) and let the solution be M^*
5: Compute the step size s from the Armijo Goldstein Equations.
6: Update M as M_{new} = M + $s(M^* - M)$
7: Evaluate the new objective $f(M_{new})$ and the new derivative vector $\nabla f(M_{new})$ using M_{new}
8: Calculate the difference in objective value: $diff$ = $abs(f(M)$ - $f(M_{new}))$
9: Update the Hessian H using the BFGS Equations [16]
10: Update the objective value: $f(M)$ = $f(M_{new})$
11: Update the derivative vector: $\nabla f(M)$ = $\nabla f(M_{new})$
12: Update the vector M: M = M_{new}
13: **until** $diff \leq \alpha$
14: Greedily set the top k entries in M as 1 to recover the integer solution.
15: Select k points accordingly

B. BMAL for Person Recognition in the SIA

To empirically validate our algorithm in the context of the social interaction assistant, we used the VidTIMIT[4] [17] and

[4]http://conradsanderson.id.au/vidtimit/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2016.2543681, IEEE Journal of Selected Topics in Signal Processing



(a) VidTIMIT dataset (b) MBGC dataset

Fig. 4. Batch Mode Active Learning on the VidTIMIT and MBGC datasets. The proposed method outperforms the baseline algorithms.

the NIST Multiple Biometric Grand Challenge (MBGC)[5] [18] datasets, both of which are publicly available. These datasets contain recordings of subjects under natural conditions where there is a redundancy of information and were hence chosen to study the performance of BMAL algorithms. The discrete cosine transform (DCT) feature was used in our experiments and Gaussian Mixture Model (GMM) was used as the underlying classifier. 25 subjects were selected for this study. The initial training set contained 250 images, the unlabeled set had 2000 images and 4500 images formed the test set, spanning all the subjects. The batch size k was taken as 10 and we studied the performance on the test set as more unlabeled samples were iteratively labeled and appended to the training set. We assumed that labeling resources were available for 1000 unlabeled samples and thus the algorithms were run for 100 iterations. The proposed approach was compared with three other BMAL schemes, which are based on heuristic scores - (i) Random Sampling, where a batch of points was randomly queried from the unlabeled pool; (ii) SVM Active Learning with Angular Diversity, where a batch of points was incrementally sampled such that at each step the hyperplane induced by the selected point maximized the angle with all the hyperplanes of the already selected points, as proposed by Brinker [11]; and (iii) Uncertainty Based Ranked Selection, where the top k uncertain points are queried from the unlabeled pool, k being the batch size.

The results are depicted in Figure 4, where the x axis denotes the iteration number and the y axis denotes the accuracy on the unseen test set [19]. We note that the proposed BMAL framework performs much better than the other methods as its accuracy on the test set grows at the fastest rate. The label complexity (defined as the number of queries needed to achieve a certain accuracy) is least in case of the proposed technique. Thus, the proposed framework succeeds in selecting the most informative unlabeled samples for manual annotation and is tremendously useful in learning a good classification model with minimal human annotation effort. We note that random sampling involves the least computational overload. Thus, in systems where computational complexity is a serious concern, random sampling can be used to select unlabeled samples (with a corresponding compromise in accuracy). However, in the design of the SIA, the data annotation and model update are performed offline; thus the proposed method is a preferred alternative as it produces more accurate solutions.

[5]http://www.nist.gov/itl/iad/ig/mbgc.cfm

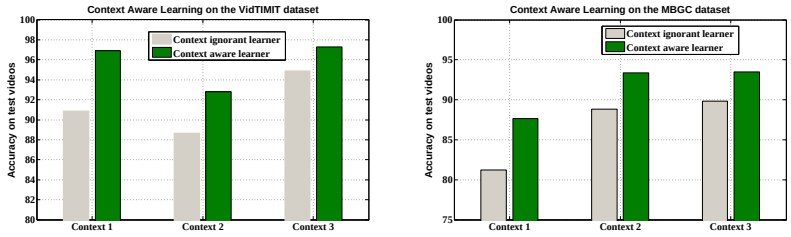
C. Person-Centered BMAL for Face Recognition

The performance of the person recognition system can be further improved by incorporating person-centeredness in the BMAL formulation. Every person has a unique daily activity schedule and expects to encounter a specific set of subjects at specific locations. For instance, in an office location, there is a higher chance of meeting work colleagues than family members; at home, the situation is reversed. Thus, person-centeredness can be integrated into the system through *contextual information*. We define context as the *location of a user*, similar to [20]. We assume that at any given location, the user is aware of the subjects to be expected in that location (for example, work acquaintances in an office setting or family members in a home setting). This was used to construct a prior probability vector depicting the chances of seeing each subject at a given location. In such a situation, the query function can be modified to ensure that the images remaining in the unlabeled video after batch selection have low entropy with respect to the subjects expected in the given context. Thus, the entropy is computed only on the subjects that are present in a given video stream:

$$f(B) = \sum_{i \in B} \rho_i - \lambda \sum_{j \in U_t - B} S^{context}(y|x_j, w^{t+1}) \qquad (9)$$

Here, $S^{context}$ is the context aware entropy term. For each unlabeled image, this term was computed from the posterior probabilities, which in turn were obtained by multiplying the likelihoods returned by the trained GMM classifier with the context aware prior. Thus, subjects not expected in a given context will have low priors and consequently, the corresponding posteriors will not contribute much in computing $S^{context}$. To simulate this situation, three contexts were arbitrarily defined and 8 random subjects (chosen from the set of 25) were assigned to each context. BMAL was used to select batches of samples from unlabeled video streams in each context. The updated classifiers were then tested on videos in the respective context. The context-ignorant learner was implemented using equal class priors in the entropy term.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2016.2543681, IEEE Journal of Selected Topics in Signal Processing



(a) VidTIMIT dataset (b) MBGC dataset

Fig. 5. Context Aware Learning on the VidTIMIT and MBGC datasets. The context-aware learner performs better than the context-ignorant learner.

Figure 5 shows the accuracies obtained on the VidTIMIT and MBGC test videos (averaged over three trials in each context) [19]. It is noted that in each context, the context aware learner produces better accuracy on test videos than the context ignorant learner. Thus, incorporation of context in the formulation further helps in querying salient images for manual annotation.

We note that, in this application, the user and the system work together and learn over time. Initially, the user supplies the chances of meeting subjects at specific locations (in the form of "yes", "no" and "maybe" inputs) for better performance of the underlying algorithm. Over time, the system adapts to the lifestyle and routine of the user and thus, the computation of the prior probability vector can be automated. This is the core idea of person-centeredness and co-adaptation.

IV. Conformal Predictions for Multi-modal Person Recognition

An essential component in any machine learning system is the computation of reliable confidence measures - which depict the system's level of certainty in its predictions. In the SIA, an incorrect system prediction about the identity of an individual can lead to an embarrassing social situation; thus, knowledge of the machine's confidence associated with each of its predictions can enable the user to react accordingly and avoid such situations. Further, it has been shown that learning from multiple sources of information is better than learning from a single source, if the sources are fused appropriately [21]. In the SIA, two sources of information - audio and video - can be combined for accurate person recognition. In this section, we detail the Conformal Predictions (CP) framework for computation of reliable confidence metrics and its application to person recognition from face and voice modalities.

A. Conformal Predictions: Background and Rationale

The theory of Conformal Predictions was developed by Vovk, Shafer and Gammerman [22], [23] based on the principles of algorithmic randomness, transductive inference and hypothesis testing. The theory is based on the relationship derived between transductive inference and Kolmogorov complexity [24] of an i.i.d (identically independently distributed) sequence of data instances. The advantage of the confidence measures obtained from the CP framework is that, the results are well-calibrated in an online setting, that is the frequency of errors, ϵ, made by the system is exactly bounded according to the confidence level, $1-\epsilon$, defined by the user. This is depicted in Figure 6, where the x axis denotes the number of test examples and the y axis denotes the cumulative errors at different confidence levels. We note that the errors are calibrated at each of the significance levels; for instance, at the 5% significance level, the number of errors committed by the system is always less than 5% of the number of test samples.

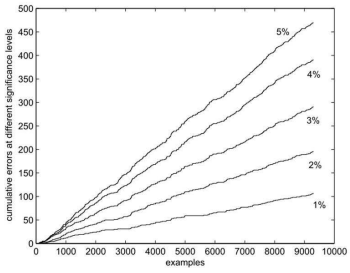


Fig. 6. Results from the Conformal Predictions Framework [22]. The errors are calibrated in an online setting.

1) Conformal Predictions in Classification: For a given test sample, say x_{n+1}, a null hypothesis is assumed that x_{n+1} belongs to the class label, say, y_p. Using this assignment, the non-conformity measures [25] of all the data points in the system so far are re-computed assuming the null hypothesis is true. Sample non-conformity measures for various classification algorithms can be found in [23]. A p-value function is defined as:

$$p(\alpha_{n+1}^{y_p}) = \frac{count\left\{i : \alpha_i^{y_p} \geq \alpha_{n+1}^{y_p}\right\}}{n+1} \tag{10}$$

where $\alpha_{n+1}^{y_p}$ is the non-conformity measure of x_{n+1}, assuming it is assigned the class label y_p. It is evident that the p-value is highest when all non-conformity measures of training data belonging to class y_p are higher than that of the new test point, x_{n+1}, which points out that x_{n+1} is *most conformal* to the class y_p. This process is repeated with the null hypothesis supporting each of the class labels, and the highest of the p-values is used to decide the actual class label assigned to x_{n+1}, thus providing a transductive inferential procedure for classification. If p_j and p_k are the two highest p-values obtained (in respective order), then p_j is called the *credibility*

of the decision, and $1-p_k$ is the *confidence* of the classifier in the decision. The conformal prediction regions are presented as regions representing a specified confidence level, Γ_ϵ, which contain all the class labels with a p-value greater than $1-\epsilon$. These regions are *conformal* i.e. the confidence threshold, $1-\epsilon$ directly translates to the frequency of errors, ϵ in the online setting [22]. The framework is summarized in Algorithm 2.

Algorithm 2 Conformal Predictors for Classification

Require: Training set $T = \{(x_1, y_1), ..., (x_n, y_n)\}$, $x_i \in X$, number of classes M, $y_i \in Y = y_1, y_2, \ldots, y_M$, classifier C

1: Get new unlabeled example x_{n+1}.
2: **for** all class labels, y_j, where $j = 1, \ldots, M$ **do**
3: Assign label y_j to x_{n+1}.
4: Update the classifier C, with $T \cup \{x_{n+1}, y_j\}$.
5: Compute non-conformity measure value, $\alpha_i^{y_j} \forall i = 1, \ldots, n+1$ to compute the p-value, P_j, w.r.t. class y_j (Equation 10) using the conformal predictions framework.
6: **end for**
7: Output the conformal prediction regions $\Gamma_{1-\epsilon} = \{y_j : P_j > 1-\epsilon, y_j \in Y\}$, where $1-\epsilon$ is the confidence level.

2) Conformal Predictors for Information Fusion: Our proposed methodology for obtaining conformal predictions in information fusion settings is fundamentally premised on multiple hypothesis testing [26]. We propose that each modality (or data feature) considered for fusion can be formulated as an independent hypothesis test, and the p-values obtained from each hypothesis test can be combined using established statistical methods [27]. Given a new test data instance, the CP framework outputs a p-value for every class label. When there are multiple data sources describing a single class label entity (e.g. different modalities like face and speech, or different image feature spaces obtained from a single face image for person recognition), we use a classifier for each individual data source with appropriate non-conformity measures, and obtain p-values for each class label uniquely for each data source. Thus, for every class label y_j, $j \in \{1, \ldots, M\}$, we have an individual null hypothesis for each data source, $H_{01}, H_{02}, \ldots, H_{0N}$, and an individual alternate hypothesis, $H_{A1}, H_{A2}, \ldots, H_{AN}$, where M is the number of class labels and N is the number of data sources. Thus, for every class label y_j, we obtain N p-values, p_i, $i = 1, \ldots, N$ one for each modality. These p-values are then combined into a new test statistic $\phi = \phi(p_1, \ldots, p_N)$, which is used to test the combined null hypothesis H_0 for class label y_j. The conformal prediction region at a specified confidence level, Γ_ϵ, is then presented as a set containing all the class labels with a p-value greater than $1-\epsilon$. The pseudo-code is depicted in Algorithm 3.

Multiple hypothesis testing has been extensively studied and a variety of methods have been proposed to combine p-values from multiple hypothesis tests. We use the Quantile combination method, where a relevant parametric cumulative distribution function (CDF), F is selected and the p-values p_is, are transformed into distributional quantiles, $q_i = F^{-1}(p_i)$ where $i = 1, 2, \ldots, k$ for each of the class labels. These q_is are subsequently combined as $\phi = \sum_i q_i$, and the p-value of the combined test H_0 is computed from the sampling distribution of ϕ. Examples of CDFs used in these methods include chi-square [28], standard normal [29], uniform [30] and logistic [31]. We use the standard normal fusion (SNF) technique for its wide usage and promising empirical performance.

Algorithm 3 Conformal Predictors for Information Fusion

Require: Number of data sources N; Training sets for each data source $T_1 = \{(x_{11}, y_1), \ldots, (x_{1n}, y_n)\}, \ldots, T_N = \{(x_{N1}, y_1), \ldots, (x_{Nn}, y_n)\}$; Number of classes M, $y_i \in Y = y_1, y_2, \ldots, y_M$; classifiers $C_1, \ldots, C_N$ for each data source

1: Get the new unlabeled example w.r.t each data source $x_{1,n+1}, \ldots, x_{N,n+1}$.
2: Using Algorithm 2 and classifiers $C_1, \ldots, C_N$ corresponding to each data source, compute p-values p_{ij}, where $i = 1, \ldots, N$ and $j = 1, \ldots, M$.
3: **for** each class label, $y_j, j = 1, \ldots, M$ **do**
4: Compute p-value, $\hat{p_j}$ of combined hypothesis using N modalities.
5: **end for**
6: Output the conformal prediction regions $\Gamma_{1-\epsilon} = \{y_j : \hat{p_j} > 1-\epsilon, y_j \in Y\}$, where $1-\epsilon$ is the confidence level.

B. Conformal Predictions for Person Recognition in the SIA

To empirically validate our algorithm in the context of the SIA, we used the VidTIMIT [17] and the MOBIO[6] [32] biometric datasets. Both these datasets contain frontal images and speech data of subjects under natural conditions. For the video modality, the faces were automatically detected and cropped from the videos; DCT feature was extracted followed by PCA and SVM was used as the underlying classification model. The Lagrangian multipliers obtained while training an SVM were used as the non-conformity measures, as suggested in [23]. The speech signal was downsampled to 8 KHz and a short-time 256-pt Fourier analysis is performed on a 25ms Hamming window (10ms frame rate). The magnitude spectrum was transformed to a vector of Mel-Frequency Cepstral Coefficients (MFCCs). A gender-dependent 512-mixture GMM Universal Background Model was initialised using k-means algorithm and then trained by estimating the GMM parameters via the Expectation Maximization algorithm. To adapt this to the CP framework, the negative of the likelihood values generated by the GMM were used as the non-conformity scores, as suggested by Vovk *et al*. [23].

The fusion results are reported in Table I [33]. It depicts the percentage of errors made by the system at different values of the confidence level. It is evident that the results are statistically calibrated across both datasets, that is the the error rate, ϵ is consistently close to the confidence level, $1-\epsilon$. This

[6]https://www.idiap.ch/dataset/mobio

TABLE I
STANDARD NORMAL FUSION RESULTS ON THE VIDTIMIT AND MOBIO DATASETS. THE ERROR RATE IS CALIBRATED AT EACH OF THE CONFIDENCE LEVELS FOR BOTH DATASETS.

Dataset	**Percentage of Errors at Confidence Level**						
	50%	**60%**	**70%**	**80%**	**90%**	**95%**	**99%**
VidTIMIT	44.46%	35.37%	25.79%	14.91%	2.59%	0.82%	0.80%
MOBIO	46.05%	37.73%	28.92%	20.49%	7.92%	2.18%	0.91%

corroborates the usefulness of the CP framework for reliable person recognition in the SIA application.

C. Person-Centered Recognition using the CP Framework

From Algorithm 3, we note that the output of the CP framework is a set of classes where the p-value is greater than $1-\epsilon$. This may sometimes result in multiple predictions, where a particular test sample may be assigned to more than one class. This occurs mostly for high confidence predictions, where the system has to deliver accurate outputs to ensure the calibration property. For low confidence predictions, the number of multiple predictions is much lower (this has been validated by our experiments). The concept of person-centeredness can be judiciously used to handle the multiple prediction problem in the context of the SIA.

As noted before, individuals who are blind or visually impaired fall on a continuum with each person residing at a specific location based on the level of blindness. Consider the case of a low-vision individual, who can rely on his or her sensory abilities to some extent. Such a person can use a moderate to high threshold on the percentage of errors $(50-60\%)$ and thus get relatively imperfect results, but with much lower number of multiple predictions. He or she can use his or her limited vision together with the system results and come to a decision about the identity of an interaction partner. Over time however, it is possible that his or her condition may deteriorate and he or she may move towards complete blindness. In such an event, he or she becomes increasingly reliant on the machine and needs to use a much lower threshold on the percentage of errors $(5-10\%)$ to ensure perfect results from the system. Multiple predictions can be handled based on the context (a particular subject may have a very low probability of being present at a given location e.g. a home acquaintance in an office setting). The machine also adapts to the user's condition over time and can automatically decide the error threshold to avoid embarrassing social situations.

As another example, consider the case of an individual who is blind. The error threshold can be decided based on his or her location at a given point of time. For instance, when he or she is in an office meeting, a low error threshold should be selected as it is imperative to accurately identify all the subjects in the meeting. When the person is at home, a relatively higher error threshold can be selected (minimizing the number of multiple predictions), as incorrect recognition does not have severe consequences in a home setting. The threshold selection can be automated over time, based on the user's daily activity schedule. In both these instances, the user and the system work hand-in-hand and adapt over time - another example of person-centeredness and co-adaptation.

V. TOPIC MODELS FOR FACIAL EXPRESSION RECOGNITION

Understanding the emotional state of an interaction partner is crucial in initiating a conversation. Facial expression and emotion recognition are thus core components of the SIA system. Human emotional state can be described using seven discrete states (anger, disgust, contempt, fear, joy, sadness and surprise). There has been an extensive body of work explaining the seven discrete states in terms of facial Action Units (AUs), as enunciated in the Facial Action Coding System (FACS) of Ekman and Friensen [34]. While there has been substantial research in the identification of these seven states, there has been a growing need to explore a more complex space of emotions necessitating facial descriptors that can capture a richer space of emotions. To this end, we have used Latent Facial Topics (LFTs), which are atomic facial descriptors that are automatically discovered using probabilistic topic models. Our extensive empirical studies emphasize the usefulness of LFTs for discrete expression recognition and also establish that facial topics derived using this method have semantic validity and are visualizable.

Topic models were originally proposed in document analysis to extract latent topics from text documents. Recently, topic models have found increasing use in image and video analysis for tasks such as object recognition, scene segmentation, image annotation etc. [35]. The fundamental idea of topic models is that every document contains latent concepts called topics. In text mining, a topic is defined as a collection of words that can co-occur in a given corpus of documents. In our research, we adapt topic models to facial video processing, where we can either model an entire video or an individual video frame as a document and explain the activity within each facial video or image document using latent facial topics (LFTs). The facial documents are created by quantizing facial features (eg. shape or appearance features, which we call base features) obtained from face images. We used two popular probabilistic topic models, Latent Dirichlet Allocation (LDA) [36] which is unsupervised and Supervised Latent Dirichlet Allocation (SLDA) [37] which is supervised, to extract LFTs from these facial documents.

A. Facial Expression Recognition in the SIA

To study the usefulness of latent facial topics for emotion recognition in the context of the social interaction assistant, we used the Cohn-Kanade Plus (CK+) [7] database [38], which contains 327 image sequences, annotated with 7 facial expressions and 34 AUs from 118 subjects. We used subject-based Leave-One-Out strategy to evaluate the performance.

[7] http://www.pitt.edu/ emotion/ck-spread.htm

TABLE II
ACCURACIES FOR DISCRETE EMOTION RECOGNITION ON CK+ DATASET. LFTS OUTPERFORM SPTS FEATURES BY A LARGE MARGIN.

Classifier	SPTS-SVML	LFT-SVML	LFT-SVMR
Accuracy	66.68%	**85.62%**	84.4%

Using the LFT distributions as features, two classifiers were trained- Support Vector Machines using Linear (LFT-SVML) and RBF kernels (LFT-SVMR). The proposed approach was compared against the similarity-normalized shape (SPTS) features (SPTS-SVML) which were originally proposed for this database [38]. The results are depicted in Table II and corroborate that LFTs lead to improved recognition performance than SPTS features [39].

Latent Facial Topics discovered automatically using LDAs are shown in Figure 7 [39]. LFTs are probability distributions over facial words, where the probability values are presented by the length of the blue lines; we note that LFTs are implicitly meaningful, as visually evident in Figure 7. A deeper analysis also reveals a relationship with facial AUs, as defined by Ekman and Friensen [34]. For example, LFT 8 and 18 together correspond to AU 1 (Inner Brow Raiser) and LFT 49 corresponds to AU 12 (Lip corner pull). These results demonstrate the efficacy of LFTs as salient features in human emotion recognition and their semantic interpretation and visualizibility.

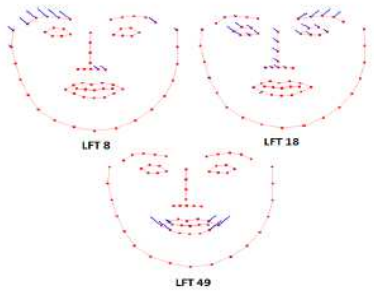


Fig. 7. Plots of LDA based LFTs from CK+ dataset. LFTs have semantic meaning and are visualizable.

B. Person-Centered Facial Expression Recognition

As evident from the above discussions, the results of facial expression/emotion recognition can be presented at different granularities. A system can be programmed to convey merely the discrete emotion or more subtle information such as the movement of the facial muscles (like an eyebrow raise, lip-corner pull etc.). Depending on a situation, a particular output mechanism may be preferred over the other. Consider the application of an individual who is visually impaired, attending an interview or in a meeting. In this case, it is imperative to gauge the mindset of the interaction partner to better understand his or her reactions. A system delivering the minute details about the facial muscle movements is immensely useful in such a scenario. On the other hand, in a more casual setting like home, conveying the intricate details may result in an unnecessary overload of information; a device furnishing merely the discrete emotion is more convenient. Such inputs may be passed manually to the system initially; over time, the system adapts to the schedule and habits of a particular user and can operate autonomously to deliver appropriate outputs. Thus, the machine and the user adapt to each other over time to solve a challenging computer vision task. This aligns with the core principles of person-centered multimedia computing.

VI. CONCLUSION AND DISCUSSION

In this paper, we presented the social interaction assistant (SIA) - an assistive technology to enrich the social interaction experience of individuals with visual impairments. We adopted a person-centered approach to the SIA so that our solutions can cater to the needs of an individual user's behavior, preferences and expectations; our user studies have shown that integration of person-centeredness is pivotal to addressing individual users' needs. While the SIA was originally designed for individuals with visual impairment, this effort naturally generalizes to the broader population. In general, the explicit needs of individuals with disabilities may be the unspecified implicit needs of the general population. For example, the SIA could be helpful for sighted individuals in remote audio communication, where the visual modality may not be available or accessible. Providing access to non-verbal visual social cues can be quite valuable in these situations. At CUbiC, we have embarked on a range of projects to serve the needs of individuals with disabilities by employing a person centered approach to the design, development and deployment of multimedia computing solutions. We believe that the PCMC approach will increasingly become the methodology of choice in the design of new technologies and applications for the general population.

REFERENCES

[1] M. Knapp, "Nonverbal communication in human interaction," in *Holt, Rinehart and Winston, New York*, 1978.
[2] P. Borkenau, N. Mauer, R. Riemann, F. Spinath, and A. Angleitner, "Thin slices of behavior as cues of personality and intelligence," in *Journal of personality and social psychology*, 2004.
[3] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis," in *Psychological Bulletin*, 1992.
[4] D. Norman, "Human-centered design considered harmful," in *Interactions - Ambient intelligence: exploring our living environment*, 2005.
[5] S. Panchanathan, T. McDaniel, and V. Balasubramanian, "Person-centered accessible technologies: Improved usability and adaptation through inspirations from disability research," in *ACM workshop on User experience in e-learning and augmented technologies in education*, 2012.
[6] ——, "An interdisciplinary approach to the design, development and deployment of person-centered accessible technologies," in *International Conference on Recent Trends in Information Technology (ICRTIT)*, 2013.
[7] S. Panchanathan and T. McDaniel, "Person-centered accessible technologies and computing solutions through interdisciplinary and integrated perspectives from disability research," in *Universal Access in the Information Society*, 2014.
[8] S. Krishna, D. Colbry, J. Black, V. Balasubramanian, and S. Panchanathan, "A systematic requirements analysis and development of an assistive device to enhance the social interaction of people who are blind or visually impaired," in *Workshop on Computer Vision Applications for the Visually Impaired*, 2008.
[9] P. Viola and M. Jones, "Robust real-time face detection," in *International Journal of Computer Vision (IJCV)*, 2004.

[10] H. Burton, A. Snyder, T. Conturo, E. Akbudak, J. Ollinger, and M. Raichle, "Adaptive changes in early and late blind: A fMRI study of braille reading," in *Journal of Neurophysiology*, 2002.
[11] K. Brinker, "Incorporating diversity in active learning with support vector machines," in *In International Conference on Machine Learning (ICML)*, 2003.
[12] S. Hoi, R. Jin, and M. Lyu, "Large-scale text categorization by batch mode active learning," in *International Conference on World Wide Web (WWW)*, 2006.
[13] Y. Guo and D. Schuurmans, "Discriminative batch mode active learning," in *Neural Information Processing Systems (NIPS)*, 2007.
[14] S. Chakraborty, V. Balasubramanian, and S. Panchanathan, "Dynamic batch mode active learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.
[15] S. Chakraborty, V. Balasubramanian, Q. Sun, S. Panchanathan, and J. Ye, "Active batch selection via convex relaxations with guaranteed solution bounds," in *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2015.
[16] J. Nocedal and S. J. Wright, *Numerical optimization*. Springer, 1999.
[17] C. Sanderson, *Biometric Person Recognition: Face, Speech and Fusion*. VDM Verlag, 2008.
[18] M. Tistarelli and M. Nixon, "Advances in biometrics, third international conference on biometrics," in *SpringerLink*, 2009.
[19] S. Chakraborty, V. Balasubramanian, and S. Panchanathan, "Generalized batch mode active learning for face-based biometric recognition," in *Pattern Recognition Journal*, 2013.
[20] A. K. Dey, G. D. Abowd, and D. Salber, "A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications," in *Human-Computer Interaction*, 2001.
[21] K. Crammer, M. Kearns, and J. Wortman, "Learning from multiple sources," in *Journal of Machine Learning Research*, 2008.
[22] G. Shafer and V. Vovk, "A tutorial on conformal prediction," in *Journal of Machine Learning Research (JMLR)*, 2008.
[23] V. Vovk, A. Gammerman, and G. Shafer, "Algorithmic learning in a random world," in *Springer-Verlag New York*, 2005.
[24] M. Li and P. Vitanyi, "An introduction to kolmogorov complexity and its applications," in *Springer-Verlag New York*, 1997.
[25] K. Proedrou, I. Nouretdinov, V. Vovk, and A. Gammerman, "Transductive confidence machines for pattern recognition," in *European Conference on Machine Learning (ECML)*, 2002.
[26] J. Shaffer, "Multiple hypothesis testing," in *Annual Review of Psychology*, 1995.
[27] T. Loughin, "A systematic comparison of methods for combining p-values from independent tests," in *Computational Statistics and Data Analysis*, 2004.
[28] R. Fisher, "Statistical methods for research workers," in *Macmillan Pub Co*, 1970.
[29] T. Liptak, "On the combination of independent tests," in *Magyar Tud. Akad. Mat. Kutato Int. Kozl.*, 1958.
[30] E. Edgington, "An additive method for combining probability values from independent experiments," in *The Journal of Psychology: Interdisciplinary and Applied*, 1972.
[31] G. Mudholkar and E. George, "The logit method for combining probabilities," in *The Journal of Psychology: Interdisciplinary and Applied*, 1979.
[32] S. Marcel, C.McCool, S.Chakraborty, V. Balasubramanian, and S. P. et al., "Mobile biometry (mobio) face and speaker verification evaluation," in *International Conference on Pattern Recognition (ICPR)*, 2010.
[33] V. Balasubramanian, S.Chakraborty, and S. Panchanathan, "Conformal predictions for information fusion: A comparative study of p-value combination methods," in *Annals of Mathematics and Artificial Intelligence (AMAI)*, 2013.
[34] P. Ekman and W. Friensen, "The facial action coding system (facs): A technique for the measurement of facial action," in *Consulting Psychologists Press*, 1978.
[35] J. Varadarajan and J. Odobez, "Topic models for scene analysis and abnormality detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.
[36] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," in *Journal of Machine Learning Research (JMLR)*, 2003.
[37] W. Chong, D. Blei, and F. Li, "Simultaneous image classification and annotation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
[38] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *IEEE Conference on Computer Vision and Pattern Recognition workshops (CVPRW)*, 2010.
[39] P. Lade, V. Balasubramanian, and S. Panchanathan, "Latent facial topics for affect analysis," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2013.

Sethuraman Panchanathan leads the knowledge enterprise development at Arizona State University, which advances research, innovation, strategic partnerships, entrepreneurship, global and economic development at ASU. Panchanathan was the founding director of the School of Computing and Informatics and was instrumental in founding the Biomedical Informatics Department at ASU. He also served as the chair of the Computer Science and Engineering Department. He founded the Center for Cognitive Ubiquitous Computing (CUbiC) at ASU. CUbiC's flagship project iCARE, for individuals who are blind and visually impaired, won the Governor's Innovator of the Year-Academia Award in November 2004. In 2014, Panchanathan was appointed by President Barack Obama to the U.S. National Science Board (NSB). He has also been appointed by U.S. Secretary of Commerce Penny Pritzker to the National Advisory Council on Innovation and Entrepreneurship (NACIE). Panchanathan is a Fellow of the National Academy of Inventors (NAI), and a Fellow of the Canadian Academy of Engineering. He is also a Fellow of the Institute of Electrical and Electronics Engineers (IEEE), and the Society of Optical Engineering (SPIE). He is currently serving as the Chair-Elect in the Council on Research (CoR) within the Association of Public and Land-grant Universities (APLU). Panchanathan's research interests are in the areas of human-centered multimedia computing, haptic user interfaces, person-centered tools and ubiquitous computing technologies for enhancing the quality of life for individuals with disabilities, machine learning for multimedia applications, medical image processing, and media processor designs. Panchanathan has published over 425 papers in refereed journals and conferences.

Shayok Chakraborty is an Assistant Research Professor in the School of Computing, Informatics and Decision Systems Engineering (CIDSE) at Arizona State University. He is also an Associate Director of ASU's Center for Cognitive Ubiquitous Computing (CUbiC) research center. He received his PhD in Computer Science from ASU in 2013 under the mentorship of Dr. Sethuraman Panchanathan. He has worked as a Post-doctoral researcher at Intel Labs, Oregon and in the Electrical and Computer Engineering department at Carnegie Mellon University. Shayok's research interests include computer vision, machine learning and assistive technology. He has actively published his work in premier conferences and journals in these areas. He is a member of IEEE.

Troy McDaniel is an Assistant Research Professor in the School of Computing, Informatics, and Decision Systems Engineering at Arizona State University. He is also an Associate Director of ASU's Center for Cognitive Ubiquitous Computing, and the Research Director of ASU's IGERT program, Alliance for Person-centered Accessible Technologies. Dr. McDaniel's research interests include haptics, human-computer interaction, assistive technologies and rehabilitative technologies. For over a decade, he has explored how our sense of touch can be better utilized by technology as a communication channel. He has over 30 peer-reviewed papers in premier haptics and human-computer interaction conferences and journals.