XXI ENMC
ENCONTRO NACIONAL DE MODELAGEM COMPUTACIONAL
IX ECTM
ENCONTRO DE CIÊNCIA E TECNOLOGIA DE MATERIAIS

08 a 11 de Outubro de 2018
Instituto Federal Fluminense
Búzios - RJ

MODELO COMPUTACIONAL DE DETECÇÃO E RECONHECIMENTO DA LINGUAGEM BRASILEIRA DE SINAIS

Marcos Batista Figueredo[1] - mbfigueredo@uneb.br
[1]Universidade do Estado da Bahia, Departamento de Ciências Exatas e da Terra II - Alagoinhas, BA, Brasil

Resumo. *O objetivo desta pesquisa é a apresentação de um modelo capaz de realizar o reconhecimento de gestos, onde a ideia e que estes modelo seja capaz de realizar a tradução tanto da linguagem de sinais para o professor como vice e versa. Este processo vai extrair características do movimento e posição das mãos e braços, a partir de imagens dinâmicas capturadas de uma câmera e identificar padrões de sinais utilizados pelos usuários de Libras (Língua Brasileira de Sinais). O aplicativo será capaz de capturar as imagens e, com a utilização de filtros específicos de processamento de imagens, identificar as características de cada gesto realizado. Os dados identificados são comparados a padrões pré-estabelecidos no software com o objetivo de identificar os sinais e reproduzir em forma de texto ou áudio as letras identificadas. O grande desafio e foco dessa pesquisa será a criação de um mecanismo preciso para comparação desses dados, pois com a variedade de informações extraídas o processo pode se tornar lento e apresentar uma baixa taxa de rastreamento. Este trabalho tem suas contribuições centradas no eixo social e tecnológico.*

Palavras Chave: *Detecção Gestos, LIBRAS, Redes Neurais, Visão Computacional*

Abstract. *The objective of this research is to present a model capable of of realizing gesture recognition, where the idea is that these models be able to perform the translation of both sign language to the teacher and vice versa. This process will extract characteristics of the movement and position of hands and arms, from dynamic images captured from a camera and identify patterns of signals used by users of Libras (Brazilian Sign Language). The application will be able to capture the images and, with the use of specific image processing filters, identify the characteristics of each gesture performed. The identified data are compared to pre-established standards in the software in order to identify the signals and reproduce in the form of text or audio the letters identified. The great challenge and focus of this research will be the creation of an accurate mechanism for comparing this data, because with the variety of information extracted the process can become slow and display a low crawl rate. This work has its contributions centered in the social and technological axis.*

Keywords: *Detection Gestures, LIBRAS, Neural Networks, Computational Vision*

1. INTRODUÇÃO

As mudanças constantes no campo educacional, tem buscado fomentar e aumentar o olhar para a diversidade de quem aprende. Sempre com respeito à identidade e à diferença, buscando promover a pluralidade cultural, racial e portadores de deficiências seja por meio de leis ou regulamentações.

Para alunos com necessidades especiais, em particular os surdos e mudos que somam cerca de 9.7 milhões de pessoas [BRASIL, 2011], terem direito ao ensino aprendizagem de forma igualitária, é necessária a presença do intérprete da Linguagem Brasileira de Sinais em parceria com o professor da disciplina [BRASIL, 2008], [CARNIEL, 2018].

Essa dificuldade em classe também pode ser estendida a outros ambientes comuns aos ouvintes, no âmbito da sociedade, como por exemplo, um seminário, uma palestra, onde não tenha a presença de intérprete da língua de sinais, além de hospitais onde o problema é ainda mais graves.

A formação do professor para o ensino de alunos com deficiência sempre esta em pauta nas discussões educacionais sobre inclusão. Aliando ao fato de que o professor deve incorporar, em sua prática, esta linguagem. Porem, no ensino superior enfrenta diversas barreiras, entre elas a falta de interpretes e a não formação da maioria dos professores na linguagem brasileira de sinais [DA SILVA et al., 2018].

A Linguagem Brasileira de Sinais (Libras) foi estabelecida na Lei n°10.436/2002[BRASIL, 2002], como língua oficial das pessoas surdas e de acordo com o próprio termo, a Libras é utilizada somente no Brasil e representa uma forma de comunicação e expressão, em que o sistema linguístico é de natureza visual-motora, com estrutura gramatical própria.

Em contrapartida a interação com os dispositivos de computação avançou de tal forma que diversos dispositivos já são capazes de compreender uma linguagem e traduzi-la. Está tecnologia vem se incorporando em nossas vidas e a utilizamos em trabalhos, compras, comunicação e até mesmo no entretenimento.

As tecnologias na computação, comunicação e exibição progridem ainda mais, mas as técnicas existentes podem se tornar um gargalo na utilização efetiva do fluxo de informações disponíveis. Para usá-los eficientemente, a maioria dos aplicativos de computador exige mais e mais interação. Por essa razão, a interação humano-computador (IHC) tem sido um campo de pesquisa ativo nos últimos anos[PREECE et al., 2015], [BEACKER, 2014],[HELANDER, 2014].

A detecção e reconhecimento de sinais tem se tornado natural a diversos dispositivos de *interface* para interação entre computadores e humanos. Usar as mãos como um dispositivo tem ajudado as pessoas a se comunicarem com os computadores de uma maneira mais intuitiva.

Os sinais e movimentos dos membros superiores assim representam o meio de comunicação não verbal, variando de ações simples (alfabeto e sistema de numeração) até outras mais complexas (como expressar sentimentos ou se comunicar com os outros). A detecção e o reconhecimento sinais é um termo que refere-se coletivamente a todo o processo de rastreamento de gestos humanos a sua representação e conversão a comandos semanticamente significativos.

Este trabalho propõe uma abordagem computacional para a tradução de Libras para o português e vice versa ,para uso em sala de aula, a partir de imagens obtidas por câmeras e recursos de visão computacional. Sendo desenvolvido pelo grupo de pesquisa em Modelagem e Simulação de Biossistemas da Universidade do Estado da Bahia e conta com professores,inclusive de Libras, alunos da graduação e pós-graduação da instituição.

Este artigo está organizado em seções. A primeira seção é essa introdução, a seção 2 ap-

resenta. o objeto a ser modelado e seus desafios, na seção 3. Apresentamos as informações relacionadas ao desenvolvimento da pesquisa como a revisão de literatura, o problema abordado, a solução proposta e implementada são mostradas na seção 4. A forma de abordar os experimentos, os resultados e as discussões são descritos na seção 5. Por fim, as considerações finais são apresentadas na seção 6.

2. O OBJETO A SER MODELADO

Em todas as línguas de sinais [BRAGA, 2017], inclusive na Libras, cada palavra é representada por um sinal, por isso é incorreto caracterizar os sinais da Libras como simples gestos ou mímicas, uma vez que se diferem por regras gramaticais específicas [BRASIL, 2002]. As línguas de sinais são chamadas de gestual-visual porque o responsável para emitir a comunicação são as mãos por meio dos sinais, e o receptor são os olhos.

A Libras é direcionada para pessoas surdas, surdo-cegas e até mesmo para pessoas surdas que não possuem braços. As pessoas surdas "escutam" com os olhos, através dos sinais direcionados a elas. Já as pessoas surdo-cegas usam o toque para "ouvir", elas seguram as mãos do emissor (pessoa que faz os sinais) para entender o que está sendo dito. As pessoas surdas que não possuem braços/mãos fazem sinais com os pés, porém os sinais são adaptados para esse tipo de comunicação [RODRIGUES and DE QUADROS, 2015].

Nesta língua, utiliza-se do próprio corpo e seus movimentos de forma muito expressiva, demostrando atitudes, comportamentos e sentimentos dos mais diversos tipos.

Com relação à gramática da Língua de Sinais, deve-se ressaltar a estrutura frasal particular dessa linguagem [GEDIEL et al., 2016]. Assim, enquanto na Língua Portuguesa, usa-se uma sequência sujeito→verbo→objeto, na Libras, usa-se objeto→verbo→sujeito ou objeto→sujeito→verbo. Essa estrutura diferenciada se baseia no conceito de que para os surdos o "objeto" da frase vem sempre antes do verbo ou sujeito, dando sentido ao que é dito.

Quando se analisa os níveis fonológicos e morfológicos da Língua Brasileira de Sinais podemos apontar alguns parâmetros que constituem cada sinal: os sinais são formados a partir da combinação do movimento das mãos com um determinado formato e em um determinado lugar, onde esse lugar pode ser uma parte do corpo ou um espaço em frente ao corpo.

Temos ainda a Configuração de Mão, por exemplo a Figura 1, para as formas das mãos, que podem ser da datilologia (alfabeto manual) ou outras formas com significados próprios. Alguns gestos possuem a mesma configuração de mão, porém diferenciam entre si devido a outros parâmetros, tais como movimentação e localização. Os Ponto de Articulação que são os lugares onde incide a mão configurada, podendo estar tocando alguma parte do corpo ou estar em um espaço neutro que vai do meio do corpo até acima da cabeça.

Os sinais podem ter movimento ou não e caso tenham podem circular o corpo ou serem isolados durante o sinal, o que muda completamente seu significado. A Orientação desse movimento é a direção em que é feito. Normalmente a sua inversão significa a ideia de oposição ou contrário. Caso o gesto não tenha movimento não faz sentido se falar em orientação [CAPOVILAO and RAPHAEL, 2004].

Portanto o objeto a ser modelado são sinais gestuais de Libras, capturados com uso de câmeras, para um conjunto específico de palavras comumente utilizadas em classes universitárias. O método tem como premissa o uso de redes neurais artificiais e classificadores em cascata utilizando em vídeos de referência dos sinais que representam uma palavra específica. Para tal, espera-se definir um conjunto mínimo de sinais e permitir que o modelo evolua ao

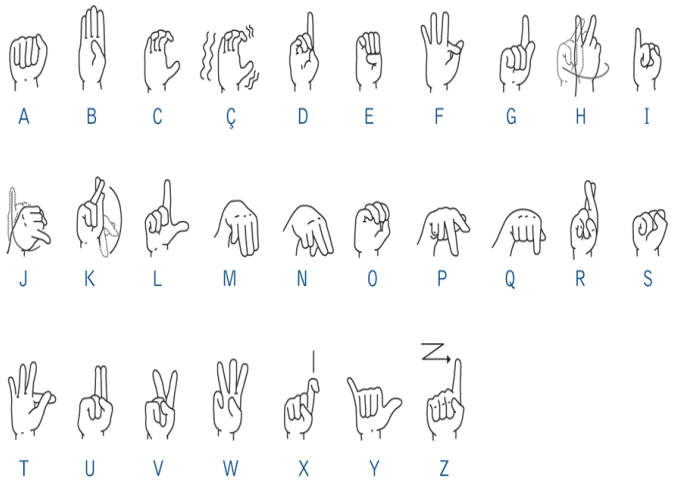

Figure 1- Letras e sua simbologia no alfabeto de LIBRAS
Fonte: baseado em [CAPOVILAO and RAPHAEL, 2004]

longo do tempo.

Serão passos importantes para a aplicação do modelo a avaliação e extração de características com utilização das informações de esqueleto e rastreamento das mãos; configurar e aplicar o descritor de características SURF por quadro para definição do descritor de cada sinal; construir, configurar e aplicar uma rede neural capaz de aprender com novos gestos; criar um conjunto de treinamento balanceado com as características extraídas dos vídeos; implementar um método para descrever uma assinatura que referencia cada sinal em vídeo utilizando histogramas das características; desenvolver um método de conversão do sinal de libras em audio e do audio para o sinal; definir um método de avaliação dos resultados.

O RECONHECIMENTO DE SINAIS

Um modelo computacional capaz de auxiliar a comunicação entre pessoas surdas e ouvintes, pode facilitar a inclusão social do adolescente e adulto. Detectar e reconhecer sinais do sistema de Libras depende de uma grande quantidade de informações e um processo de treinamento de classificadores para melhora de resultados. Porém, isto aumenta o custo dos sistemas e pode até inviabilizar a sua aplicação. Com o uso da base de vídeos capturados a partir de câmeras é apresentado o problema de reconhecer gestos utilizando os quadros dos vídeos de forma completa limitando apenas a definição de escala de cores, utilizando dados do movimento de toda região superior do corpo w métodos de rastreamento da mão.

O segundo problema inerente ao reconhecimento de sinais é a codificação dos vídeos em áudio de modo a manter uma quantidade de características do movimento executado que faça possível a diferenciação de cada gesto. Outro aspecto do problema abordado é a classificação baseado no conjunto de treinamento utilizando algoritmos de reconhecimento.

Dessa forma principal desafio do modelo é a detecção e reconhecimento dos sinais em um ambiente ruidoso. Como o modelo será baseado em visão computacional envolve o manuseio de um número considerável de graus de liberdade (DoF), grande variabilidade da aparência 2D dependendo do ponto de vista da câmera (mesmo para o mesmo gesto), diferentes escalas de

silhueta (resolução espacial) e muitas resoluções para a dimensão temporal (isto é, variabilidade da velocidade do gesto).

Existe diversas abordagens descritas na literatura [Nowozin et al., 2017], [Aubauer et al., 2018],[Kim et al., 2017],[Bautista et al., 2016],[Yang et al., 2018] que em geral utilizam o esquema apresentado na Figura 2

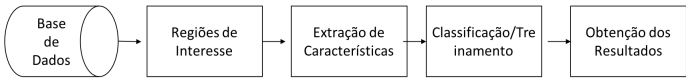

Figure 2- metodologia aplicada comumente na literatura
Fonte: Próprio autor

A primeira é onde as imagens são continuamente obtidas com uso da câmera para montagem do *dataset* de treinamento ou para a classificação em seguida destacamos a posição e forma da mão bem como também a posição do braço. Destacamos que a expressão facial também se revela importante na linguagem de libras mas nesta fase do trabalho não vamos nos dedicar a esta etapa. Nela é feito todo o tratamento da imagem com filtragem e nivelamento. Em seguida são extraídas as características principais e esse vetor é utilizado na próxima etapa para treinamento da rede e ou para classificação. Na última etapa são analisadas as quantidades de acerto positivo, falso positivo e negativos.

Neste cenário, equilibrar a compensação precisão-desempenho-utilidade de acordo com o tipo de aplicação, o custo da solução e vários critérios, como desempenho em tempo real, robustez, escalabilidade e independência do usuário é necessário. Também inclui-se no processo a sintetização de voz, visto que o professor fala e seu som é transformado em sinal para o aluno e ao gesticular sua informação é processada e passada ao professor. Este processo é realizado em tempo real e o modelo deve ser capaz de analisar tanto a imagem, na taxa de quadros do vídeo de entrada para dar ao usuário um feedback instantâneo do gesto manual reconhecido como a voz.

A robustez desempenha um papel importante no reconhecimento de sinais manuais com sucesso sob diferentes condições de iluminação e fundos desordenados. O sistema também deve ser robusto contra rotações de imagem no plano e fora do plano. A escalabilidade ajuda no manuseio de um grande vocabulário de gestos que pode ser incluído com um pequeno número de primitivos. Isso torna a composição de diferentes comandos de gestos facilmente controlados pelo usuário.

A independência do usuário cria o ambiente em que o sistema pode ser manipulado por diferentes usuários, em vez de usuários específicos, e deve também reconhecer sinais executados por humanos de diferentes tamanhos e cores.

Nesta tarefa será utilizada uma rede neural artificial (RNA) para o reconhecimento dos sinais e processos de segmentação, detecção de pele e extração de características com base no algoritmo de Viola and Jones [2004] para detecção. A RNA constitui-se de pelo menos uma camada de entrada, uma camada de saída e pelo menos uma camada escondida, podendo ter mais de uma camada escondida, se necessário. Como a RNA aprende e reconhecer padrões a partir de treinamentos utilizaremos um conjunto de dados disponível e gratuitos.

Para treinamento da rede será utilizado o dicionário online "INES" [LIRA and SOUZA, 2011] contendo 3853 sinais/itens léxicos. Cada um destes verbetes é apresentado em uma sequência de vídeo com resolução 240×180 pixels, executado por uma única mulher em ambiente controlado de fundo branco. Esse dicionário tem por finalidade ensinar LIBRAS. Para

tarefas de reconhecimento automático de palavras em LIBRAS, é necessária uma base de dados contendo mais amostras de cada palavra, em diversos cenários. Por este motivo, este trabalho propõe uma nova base de dados mais adequada para treinar e testar sistemas de reconhecimento de palavras em LIBRAS.

3. PROPOSTA ATUAL

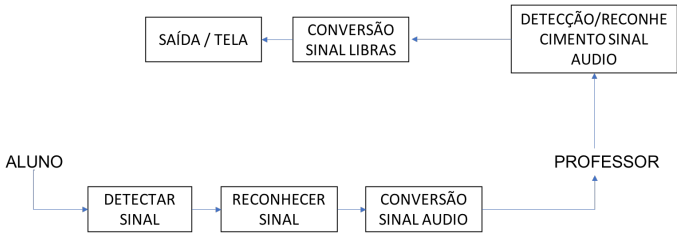

Figure 3- Proposta inicial da modelagem do problema e sua solução
Fonte: Próprio autor

O modelo de reconhecimento e tradução de LIBRAS discutido neste trabalho está ilustrado no diagrama da Figura 3. De forma circular, a modelagem detecta o discurso do professor e o converte em sinal LIBRAS, que será interpretado por um indivíduo virtual e exibido em tempo real pela saída. Neste ciclo será realizada a detecção e reconhecimento de fala [Rabiner and Juang, 1993], [Qian et al., 2016],[Toshniwal et al., 2017],[Besacier et al., 2014], para em seguida a conversão em LIBRAS.

O caminho do aluno para o professor se dará por meio da detecção do sinal vindo do aluno usando um sensor que capture os movimentos do aluno para em seguida ocorrer a interpretação deste sinal e sua conversão para audio utilizando a metodologia ASR.

4. CONCLUSÃO

NA primeira etapa uma estrutura de código aberto para reconhecimento geral de sinais será apresentada e testada com sinais isolados de linguagem de sinais. Faremos uso do Kinect, uma câmera de profundidade que torna a reconstrução 3D em tempo real facilmente aplicável. O reconhecimento é feito usando modelos ocultos de Markov com uma densidade de observação contínua. O modelo também oferecerá uma maneira fácil de inicializar e treinar novos gestos ou sinais, realizando-os várias vezes na frente da câmera.

Os testes primários apresentaram resultados com uma taxa de reconhecimento de 97% mostram que as câmeras de profundidade são adequadas para o reconhecimento da linguagem de sinais.

Este trabalho possui duas contribuições principais. A primeira delas é a disponibilização de um novo sistema de tradução, em tempo real, de LIBRAS contendo sequências de vídeo, distribuídas em verbetes.

Apesar de LIBRAS ser obrigatório em diversos cursos de graduação existe um passivo de professores que não tiveram acesso a esta linguagem assim este trabalho terá como segunda contribuição a melhoria da relação/comunicação entre pessoas com deficiência auditiva e este passivo.

Este trabalho tem como contribuição social a apresentação de uma proposta para utilização de recursos de reconhecimento de gestos aplicados à área de educacional no que tange a possibilidade de tradução simultânea de LIBRAS. Pelo aspecto tecnológico um sistema de reconhecimento para um conjunto específico de gestos é esperado como contribuição do projeto, fornecendo ao universo acadêmico uma maior entendimento sobre a comunicação entre homem máquina. Este trabalho tem como contribuição a definição de um método para classificação de sinais baseado em redes neurais, onde serão descritos:

1. A avaliação da utilização do algoritmo da rede em associado à informação espacial como descritor de quadros de vídeo capturados com o Microsoft Kinect® com a utilização da informação de esqueleto do personagem e rastreamento de partes específicas da imagem.

2. A implementação e análise de algoritmos para a conversão do sinal de vídeo em áudio e vice versa, considerando aspectos como desempenho, alinhamento e custo.

3. Definição de uma forma de representar um vídeo de profundidade em cadeia de caracteres mantendo os aspectos de movimentos em cada gesto.

Este trabalho está em fase de elaboração e planejamento e será conduzido por um professor, quatro alunos de iniciação científica e um aluno de mestrado. Este grupo pretende apresentar em 2 anos um produto possa ser utilizado no âmbito da Universidade do Estado da Bahia em seus diversos cursos e campus.

REFERÊNCIAS

Roland Aubauer, Artem Ivanov, Thomas Kandziora, and Manfred Schacht. System and method for contactless detection and recognition of gestures in a three-dimensional space, March 20 2018. US Patent 9,921,690.

M. Á. Bautista, A. Hernández-Vela, S. Escalera, L. Igual, O. Pujol, J. Moya, V. Violant, and M. T. Anguera. A gesture recognition system for detecting behavioral patterns of adhd. *IEEE Transactions on Cybernetics*, 46(1):136–147, January 2016. ISSN 2168-2267. doi: 10.1109/TCYB.2015.2396635.

Ronald M BEACKER. *Readings in Human-Computer Interaction: toward the year 2000*. Elsevier, 2014.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100, 2014.

José Luiz BRAGA. Generative communication: a dialogue with oliver sacks. *MATRIZes*, 11 (2):35–55, 2017.

BRASIL. *Lei nº 10.436 de 24 de abril de 2002, Dispõe sobre a Língua Brasileira de Sinais – Livras e dá outras providencias*. Diário Oficial da União, 2002. URL `http://www.planalto.gov.br/ccivil_03/leis/2002/l10436.htm`.

BRASIL. *Política Nacional de Educação Especial na Perspectiva da Educação Inclusiva*. Secretaria de Educação Especial, 2008. URL `http://portal.mec.gov.br/seesp/arquivos/pdf/politica.pdf`.

BRASIL. Instituto brasileiro de geografia e estatística - ibge, características da população e dos domicílios resultados do universo. *Censo Demográfico 2010*, 1:161, 2011.

Fernando César CAPOVILAO and Walkiria Duarte RAPHAEL. *Enciclopédia da língua de sinais brasileiras: o mundo do surdo em libras*, volume 8. Edusp, 2004.

FAGNER CARNIEL. A reviravolta discursiva da libras na educação superior. *Revista Brasileira de Educação*, 23:e230027, 2018.

Osni Oliveira Noberto DA SILVA, Theresinha Guimarães Miranda, and Miguel Angel Garcia Bordas. História e panorama da formação de professores de educação especial no brasil. *Revista Cocar*, 11(22):109–126, 2018.

Ana Luisa Borba GEDIEL, Charley Pereira Soares, and Cristiane Lopes Rocha de Oliveira. O ambiente virtual como aliado no processo de ensino e aprendizagem da libras. *Revista (Con) textos Linguísticos*, 10(16):24–37, 2016.

Martin G HELANDER. *Handbook of human-computer interaction*. Elsevier, 2014.

S. Y. Kim, H. G. Han, J. W. Kim, S. Lee, and T. W. Kim. A hand gesture recognition sensor using reflected impulses. *IEEE Sensors Journal*, 17(10):2975–2976, May 2017. ISSN 1530-437X. doi: 10.1109/JSEN.2017.2679220.

Guilherme de Azambuja LIRA and Tanya Amara Felipe de SOUZA. Dicionário da língua brasileira de sinais v3, 2011, 2011. URL `http://www.acessibilidadebrasil.org.br/libras_3/`.

Sebastian Nowozin, Pushmeet Kohli, and Jamie Daniel Joseph Shotton. Gesture detection and recognition, April 11 2017. US Patent 9,619,035.

Jenny PREECE, Yvonne Rogers, and Helen Sharp. *Interaction design: beyond human-computer interaction*. John Wiley & Sons, 2015.

Yanmin Qian, Tian Tan, and Dong Yu. Neural network based multi-factor aware joint training for robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(12):2231–2240, 2016.

Lawrence R Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*, volume 14. PTR Prentice Hall Englewood Cliffs, 1993.

Carlos Henrique RODRIGUES and Ronice Müller DE QUADROS. Diferenças e linguagens: a visibilidade dos ganhos surdos na atualidade. *Revista Teias*, 16(40):72–88, 2015.

Shubham Toshniwal, Tara N Sainath, Ron J Weiss, Bo Li, Pedro Moreno, Eugene Weinstein, and Kanishka Rao. Multilingual speech recognition with a single end-to-end model. *arXiv preprint arXiv:1711.01694*, 2017.

Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

A. Yang, S. M. Chun, and J. Kim. Detection and recognition of hand gesture for wearable applications in iomt. In *Proc. 20th Int. Conf. Advanced Communication Technology (ICACT)*, pages 1046–1053, February 2018. doi: 10.23919/ICACT.2018.8323932.