Reconhecimento de Ações por RNA em Aplicações de Robótica Social ⋆

Vitor T. Basilio * **Kevin B. de Carvalho** **
Alexandre S. Brandão ***

* *NERO,Departamento de Engenharia Elétrica, Universidade Federal de Viçosa, MG, (e-mail: thinassi.basilio@gmail.com).*
** *NERO,Departamento de Informática , Universidade Federal de Viçosa, MG (e-mail: kevin.carvalho@ufv.br)*
*** *NERO,Departamento de Engenharia Elétrica, Universidade Federal de Viçosa, MG, (e-mail: alexandrebrandao82@gmail.com) .*

Abstract: Action recognition has been gaining interest in research due to its vast array of applications. In this paper, it is proposed an action recognition method using Artificial Neural Networks(ANN), applying a dimension reduction to the system inputs. This method is intended to be used on social robotics, thus it required a personalized dataset which has eleven action classes, being two of them neural actions. Dynamic Time Warping (DTW) is our base line for comparison and validation of the proposed method. method, already utilized in the literature. The simulations were done applying both methods to our dataset and show the results on confusion matrixes. The proposed method has a better performance than the DTW, specially considering the neutral actions with the other ones.

Resumo: O reconhecimento de ações tem ganho espaço em pesquisas devido ao seu grande leque de aplicações. Nesse trabalho é proposto um método de reconhecimento de ações utilizando Redes Neurais Artificiais (ANN) aplicando uma redução de dimensionalidade nas entradas do sistema. O foco desse método é que ele seja utilizado para aplicações em robótica social, para tal foi necessário a criação de um banco de dados com um total de onze classes de ações, sendo duas delas ações neutras. A validação do método se dá pela comparação com o método de *Dynamic Time Warping(DTW)*, ja utilizado na literatura. As simulações são realizadas aplicando os métodos nas amostras do banco de dados e os resultados são mostrados em matrizes de confusão. Conclui-se que o método proposto tem um desempenho acima do *DTW*, em especial ao não confundir ações neutras com as outras classes de ações.

Keywords: Neural Networks; Action Recognition;Social Interactive Robotics.
Palavras-chaves: Redes Neurais;Reconhecimento de Ações;Robótica Socialmente Interativa;

1. INTRODUÇÃO

A robótica socialmente assistiva (*Social Assistive Robotics*, SAR) pode ser definida como a interseção de *Social Interactive Robotcs* (SIR) e *Assistive Robotics* (AR) (Feil-Seifer and Matarić, 2011). SIR, inicialmente definida por (Fong et al., 2003) e complementada por (Feil-Seifer and Mataric, 2005), é inspirada pela comunicação entre robô e ambiente ao seu redor, ou até mesmo entre robôs, ou seja, qualquer tipo de interação. AR é a área de estudo adequadamente definida como uma máquina que auxilia um ser humano em uma dada tarefa (Feil-Seifer and Mataric, 2005).

Diversos trabalhos têm a SAR como área de estudo, podendo ser aplicada para auxilio às pessoas idosas (Tapus et al., 2007), indivíduos em recuperação física (Kahn et al., 2001) ou até mesmo para robótica educacional (Benitti, 2012). A interação entre usuário e máquina se dá mediante diversos métodos:

a) **gestos**, através dos quais a linguagem corporal pode ser útil nos casos em que o robô e o humano interagem com o ambiente. Além disso, para terapia física o reconhecimento da posição do corpo é vital e pode realçar a comunicação em questão (Kanda et al., 2004).

b) **voz**, por ser conveniente na interação entre pessoas, pode ser também para diálogo com robôs, que utilizam gerador sintético de discurso ou voz humana pré-gravada (Eriksson, 2004).

c) **periféricos**, como *mouse*, teclados e telas *touchscreen*, uma vez que, em algumas situações, por exemplo, a utilização de gestos, como decidir pontualmente um local no mapa pode se tornar, de certa forma, tediosa. Assim, a utilização de *mouse* ou uma tela *touchscreen* pode ser mais eficiente e mais naturais para certos usuários (Huttenrauch and Eklundh, 2002) (Montemerlo et al., 2002).

⋆ Este trabalho teve suporte financeiro do CNPq,CAPES e FAPEMIG

DOI: 10.17648/sbai-2019-111602

Esse trabalho foca na comunicação dada a partir do reconhecimento de ações corporais. Ele está dividido da seguinte forma: A Seção 2 contém uma breve revisão de literatura no tema de reconhecimento de ações.A Seção 3, detalha a metodologia utilizada, incluindo a implementação da proposta e a análise comparativa. Na Seção 4, explica-se sumariamente como as simulações foram feitas para obter os resultados, que são expostos e discutidos na Seção 5. Por fim, o trabalho é concluído na Seção 6, onde também é apontada a direção dos trabalhos futuros.

2. RECONHECIMENTO DE AÇÕES

Devido às diversas aplicações, o reconhecimento de ações vem ganhando destaque no meio acadêmico nas últimas décadas: na utilização em espaços públicos para detectar atividades suspeitas, em monitoramento médico, em performance atlética e em outras atividades avançadas de interação humano-máquina, por exemplo, em robótica social (Sempena et al., 2011) (Aggarwal and Xia, 2014) (Aggarwal and Ryoo, 2011) (Moeslund et al., 2006).

Na área de processamento digital de imagens, rastreamento é equivalente ao estabelecimento de uma correspondência da estrutura de uma imagem entre consecutivos quadros baseados nas suas características principais(do inglês, *features*), ou seja, posição, velocidade, forma, textura e cor de certo objeto (Weingaertner et al., 1997).

Para determinar aspectos do corpo humano, como sua configuração e seus movimentos, é necessária a utilização de sensores, como câmeras de profundidade aliadas às técnicas de processamento de imagens, ou então dispositivos sobre a pele, como *magnetic field trackers*, luvas e roupas que utilizam rastreamento óptico, que captam aquelas informações (Mitra and Acharya, 2007).

Como o reconhecimento de ações depende da sua implementação prática, ele requer, comumente, a utilização de diferentes dispositivos de imagem e técnicas de rastreamento (Thobbi and Sheng, 2010). Nesse contexto, o presente trabalho utiliza o sensor *Microsoft Kinect v2.0* para realizar tal atividade.

Uma vez que se tenha um objeto rastreado, faz-se necessária a criação de rótulos de identificação para posterior classificação. Dentre algumas técnicas pode-se citar:

a) ***Dynamic Time Warping* (DTW)**, que compara séries temporais para encontrar o grau de similaridade entre elas. Essas séries, para o caso de reconhecimento de aões, podem ser representadas a partir de um vetor de *features* que representa as orientações e/ou posições das juntas de um esqueleto humano, ou por imagens de profundidade(Celebi et al., 2013) (Sempena et al., 2011);

b) **Espaço Vetorial**, onde o objeto de classificação pode ser representado em um espaço vetorial sem a necessidade de reconstrução 3D da estrutura. A abordagem conta com a representação de autoespaços. Tal método pode ser aplicado utilizando-se cada elemento das imagens de profundidade, ou seja, um pixel (Watanabe and Yachida, 1998);

c) ***Hidden Markov Model* (HMM)**, que realiza uma quantização da configuração de um sistema através de número finito de estados discretos, cujos valores armazenados representam uma aproximação da dinâmica do sistema. Esses estados podem ser um vetor de *features* o qual consiste na posição espacial de cada bolha caracterizadora de um objeto, bolha esta obtida ao se utilizar um rastreador de bolhas estéreo auto calibrável (Fahad et al., 2018);

d)**Redes Neurais Artificiais (RNA)**, onde diversos métodos de treinamento e otimização podem ser implementados em uma estrutura inspirada na biologia humana, onde as entradas do sistema são ponderadas pelos pesos e através de aprendizado supervisionado, são utilizadas em aplicações como classificação ou regressão(Ng and Ranganath, 2002).

Esse trabalho propõe um método de reconhecimento de ações com enfoque para aplicação em robótica social. Isso é realizado utilizando-se redes neurais simples com redução da dimensionalidade das entradas, treinadas em um banco de dados próprio. As ações são representadas através de *features*, coordenadas cartesianas das juntas do corpo, extraídas por sensores. O método é validado de forma comparativa com a técnica *Dynamic Time Warping*, utilizada em diversos trabalhos na literatura.

3. METODOLOGIA

Nesta seção, é detalhado como o banco de dados foi criado e como os diferentes métodos de reconhecimento de ações foram aplicados para posteriormente comparar seus desempenhos.

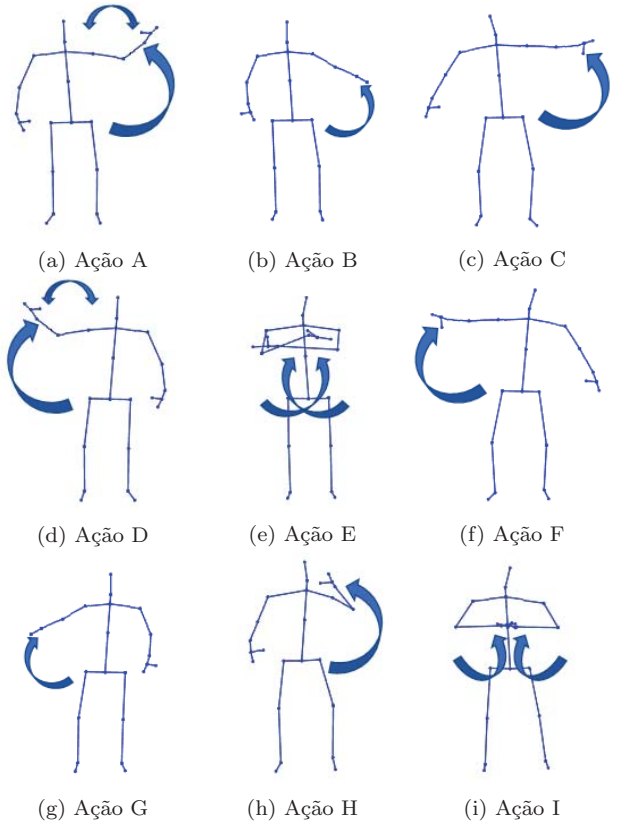
3.1 Banco de Dados

Esse trabalho de reconhecimento de ações tem seu foco na robótica social, para realização de tarefas como seguir o usuário, parar, executar uma missão e à sua origem tudo isso através de comandos de um usuário. Tendo em vista que os bancos de dados, encontrados na internet, tais como o utilizado por (Li et al., 2017), têm classes de ações incompatíveis para as tarefas pretendidas para o resultado deste trabalho. Por essa razão optou-se pela criação de um banco de dados próprio.

Isso foi feito utilizando o sensor *Kinect v2.0* para extrair a imagem de profundidade e as informações das juntas do esqueleto. Nove classes de ações padrão foram escolhidas para integrar o banco de dados, como pode ser visto na Figura 1. Todas as ações partem da posição de descanso quando se está em pé, e são um gesto de adeus com a mão direita (A); levantar o braço direito até, aproximadamente, 45 graus em relação ao tronco (B); levantar o braço direito até estar aproximadamente 90 graus com o torso (C); gesto de adeus com a mão esquerda (D); sobrepor os braços na frente do torso, formando um X (E); levantar o braço esquerdo até estar, aproximadamente, 90 graus com relação ao torso (F); levantar o braço esquerdo de até aproximadamente 45 graus com o tronco (G); levantar o braço para frente com a palma da mão direita para cima, sendo um sinal de ”vem”(H); e juntar as mãos na frente da cintura (I). Como é esperado que em robótica social o robô esteja constantemente vendo o usuário, foram adicionadas duas classes para serem identificadas como ações neutras, isto é, sem comandos atrelados a elas. A primeira ação neutra contém amostras do usuário parado, ou utilizando o celular e a segunda ação neutra contém o

DOI: 10.17648/sbai-2019-111602

usuário andando. Dessa forma, o agente poderá distinguir essa ações neutras como movimentos naturais do usuário não seguindo comando algum quando uma dessas ações for identificada.



(a) Ação A (b) Ação B (c) Ação C

(d) Ação D (e) Ação E (f) Ação F

(g) Ação G (h) Ação H (i) Ação I

Figura 1. Ações utilizadas na criação do banco de dados

Neste trabalho, cada ação foi capturada usando uma janela de 1,66 segundos, com uma taxa de captura de 30 FPS (quadros por segundo), com um total de 50 quadros. Cada quadro consiste em um grupo de *features*, as juntas do esqueleto, 25 no total, e elas são armazenadas em uma matriz de forma:

$$F_k = \begin{bmatrix} X_1 & Y_2 & Z_3 \\ \vdots & \vdots & \vdots \\ X_i & Y_i & Z_i \end{bmatrix} \quad (1)$$

onde k indica o quadro e as colunas representam as coordenadas X,Y e Z da i-ésima junta do esqueleto.

Aos olhos de um observador humano uma ação feita em diferentes posições podem ser assumidas como idênticas,o que não ocorre quando vistos pelo sensor *Kinect*, terão coordenadas diferentes para suas juntas,mesmo que pertençam a mesma classe, simplesmente por terem sido feitas em lugares distintos. Para evitar esse problema, todos os *features* são centralizados em relação às coordenadas da junta do ombro esquerdo.

Após centralizar,a ação completa é armazenada como a concatenação subsequente da matriz de *features* de todos os quadros, dado por :

$$A_n = [F_1 \dots F_k] \quad (2)$$

No caso deste trabalho, o banco de dados tem 100 amostras de cada classe, totalizando 1100 amostras no total, coletadas com pessoas diferentes para ter variabilidade de configurações de juntas distintas.

3.2 Dynamic Time Warping

Esse método consiste na comparação de séries temporais para compreender quão similares são. Esse trabalho utilizou *features* extraídos pelo sensor *Kinect* para representar cada quadro da ação. Com intuito de selecionar apenas os *features* mais relevantes para a ação, apenas 8 juntas foram utilizadas, como mostra a Figura 2, em que os ombros, cotovelos, pulsos e mãos estão destacados no esqueleto,foram escolhidas tais juntas por se tratarem das partes do corpo que mais se movem durantes a realização das ações contidas no banco de dados.

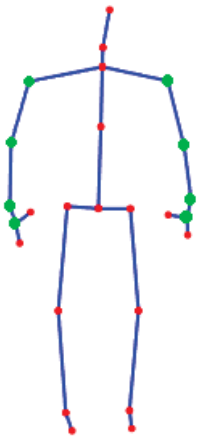


Figura 2. Juntas do corpo com destaque para as juntas utilizadas na classificação;

Foi escolhida uma série temporal para cada classe de ação para ser tomada como referência para comparação. Ela é representada por sua Matriz A_n.

O reconhecimento da ação é obtido através da comparação da série temporal da ação a ser classificada com a série temporal da referência de todas as classes. C_n contém o custo da ação A_x em relação à n-ésima ação de referência. Sabendo disso, a classificação se dará pela busca do menor custo, dada por :

$$Custo = AçãoAClassificar - AçãoReferência \quad (3)$$

O esquema de classificação utilizando o DTW pode ser encontrado no fluxograma da Figura 3.

3.3 Redes Neurais Artificiais

Similarmente ao que foi feito com o método DTW, utilizou-se apenas as 8 juntas mais relevantes para as ações na análise. A classificação de ações proposta nesse trabalho utilizando rede neural é realizada inicialmente com um pré processamento das amostras, dado por

$$A_y = \lambda\{A_x A_x^T\}, \quad (4)$$

onde $\lambda\{\cdot\}$ indica o cálculo dos autovalores da matriz $A_x A_x^T$. Note que a entrada da RNA será um vetor de 8 elementos

DOI: 10.17648/sbai-2019-111602

Figura 2. Fluxograma da classificação por DTW

Figura 3. Fluxograma da classificação por Rede Neural

representativos. A rede neural usada tem uma camada oculta com apenas dez neurônios e todas as camadas são totalmente conectadas. A função de ativação usada na camada escondida foi de tangente *sigmoid* e para a camada de classificação *softmax*. O método de treinamento utilizado foi o de Regularização Bayesiana com otimização Levenberg-Marquadt.

O treinamento da rede foi realizado utilizando 70 porcento do banco de dados como treinamento, 15 porcento como validação e 15 porcento para testes. Sendo as amostras escolhidas aleatoriamente. Esse procedimento foi repetido 10 vezes para poder ter uma média do desempenho.

O esquema de classificação utilizando o método proposto com redes neurais pode ser encontrado no fluxograma da Figura 3.

4. DISCUSSÃO E RESULTADOS

A validação do método proposto por esse trabalho foi feita através de simulações utilizando o banco de dados criado com o método proposto e o método já consolidado de DTW, tendo assim um comparativo de desempenho em relação às mesmas amostras.

Como já mencionado para o DTW, foi tomado uma série temporal do banco de dados como base para cada classe de gestos e então classificou-se todas as outras amostras.

Os resultados das simulações estão na forma de Matriz de Confusão, onde pode se saber as classes em que cada amostra foram classificadas. Vale destacar que as classes J e K são ações neutras, a primeira sendo o usuário parado e a segunda ele andando.

Os resultados das simulações com o DTW e o método proposto se encontram na Tabela 1. Onde os resultados de cima são os do DTW e os de baixo da RNA.

Inicialmente é importante ressaltar a aplicação desse trabalho. Ele propõe um método de reconhecimento de ações para ser utilizado em uma aplicação de robótica social onde as ações do usuário serão as fontes de comando para o agente. Isso implica em dois pontos importantes, o primeiro é que ações neutras devem ter um rigor maior em relação à precisão, o agente estar executando um comando e confundir o usuário andando com um ação padrão e mudar a tarefa sendo executada inviabiliza a aplicação.O

Tabela 1. Resultados das simulações do DTW e da RNA.

	A	B	C	D	E	F	G	H	I	J	K
A	99	-	-	-	-	-	-	-	-	-	-
	96,5	-	0,4	0,2	-	0,1	-	3,2	0	0,2	0,1
B	0	100	4	-	-	-	-	-	-	1	16
	-	98,9	0,4	-	-	0,9	-	-	-	2,1	-
C	-	-	93	-	-	-	-	-	-	-	-
	-	-	95,8	-	0,4	0,9	-	-	-	0,1	-
D	-	-	-	94	-	-	-	-	-	-	-
	-	-	-	99	-	-	-	-	-	0,3	0,5
E	-	-	-	-	96	-	-	-	-	-	-
	0,8	-	0,5	-	99,1	-	-	-	0,5	0,3	1,2
F	-	-	-	6	-	100	-	-	-	-	-
	0,3	-	0,6	-	-	94,8	1	-	-	1,3	-
G	-	-	2	-	-	-	100	-	-	-	-
	-	-	0,5	-	-	1,3	99	-	-	1,3	-
H	-	-	-	-	-	-	-	100	-	-	-
	2,4	-	-	-	-	-	-	96,8	-	-	-
I	-	-	-	-	-	-	-	-	62	38	1
	-	-	0,4	-	0,2	-	-	-	96,4	0,9	5,7
J	1	-	-	-	4	-	-	-	38	61	0
	-	1,1	1,2	0,1	0	19	-	-	-	93,4	1,6
K	-	-	1	-	-	-	-	-	-	-	83
	-	-	0,2	0,7	0,3	-	-	-	2,3	1,3	90,9

segundo é que a taxa de acerto deve ser ser alta. Erros de classificação das ações padrão com as ações neutras é algo menos problemático, afinal, dar um comando para agente fazer algo e ele fazer nada não é tão danoso quanto um comando ser confundido por outro ou uma ação neutra ser confundida por uma ação padrão.

Comparando os valores da Tabela, é possível observar que o método DTW confundiu cada ação padrão com menos ações padrão diferentes, a classe que confundiu com mais classes diferentes foi a C, que erroneamente foi classificada como B, G e K. Sendo que a rede neural teve confusões em mais classes diferentes.A classe C foi confundida com todas as outras, exceto D e H, apesar de ser em menor frequência.

Pode-se perceber também que a classe de ação neutra J foi confundida com a classe I com o DTW, o que é altamente indesejável para a aplicação. Já na rede neural, essa confusão é reduzida, tendo mais de 90% de acerto nas duas classes de ação neutra.

5. CONCLUSÃO E TRABALHOS FUTUROS

Nesse trabalho é proposto um método de classificação de ações utilizando Redes Neurais Artificiais baseado em *features* do esqueleto humano, com redução da dimen-

DOI: 10.17648/sbai-2019-111602

sionalidade da entrada do sistema. Com a intenção de aplicação em robótica social, um banco de dados próprio foi coletado e testado com o método proposto baseado em RNA e o DTW. Os resultados mostraram que o DTW sofre mais com classificações de ações neutras em ações padrão. Ainda que tenha uma precisão boa, o efeito dessas confusões tornaria ele inviável para essa aplicação. Já a RNA, tem uma média de acerto superior e ainda conta com menos classificações de ações neutras em ações padrão, mostrando-se mais apta para o uso prático.

Para trabalhos futuros, será feita a aplicação *online* do método proposto e integrá-lo com um robô capaz de seguir comandos como seguir, aproximar, parar , retornar para a origem, sendo de ajuda em atividades como carregar peso para o usuário.

AGRADECIMENTOS

Esse trabalho foi apoiado pelo CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico, FAPEMIG - Fundação de Amparo à Pesquisa de Minas Gerais e CAPES-Coordenação de Aperfeicoamento de Pessoal de Nível Superior.

REFERÊNCIAS

Aggarwal, J.K. and Ryoo, M.S. (2011). Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3), 16.

Aggarwal, J.K. and Xia, L. (2014). Human activity recognition from 3d data: A review. *Pattern Recognition Letters*, 48, 70–80.

Benitti, F.B.V. (2012). Exploring the educational potential of robotics in schools: A systematic review. *Computers & Education*, 58(3), 978–988.

Celebi, S., Aydin, A.S., Temiz, T.T., and Arici, T. (2013). Gesture recognition using skeleton data with weighted dynamic time warping. In *VISAPP (1)*, 620–625.

Eriksson, J. (2004). Hands-off robotics for post-stroke arm rehabilitation. *Technical Report*.

Fahad, H., Ghani Khan, M.U., Saba, T., Rehman, A., and Iqbal, S. (2018). Microscopic abnormality classification of cardiac murmurs using anfis and hmm. *Microscopy research and technique*, 81(5), 449–457.

Feil-Seifer, D. and Mataric, M.J. (2005). Defining socially assistive robotics. In *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005.*, 465–468. IEEE.

Feil-Seifer, D. and Matarić, M.J. (2011). Socially assistive robotics. *IEEE Robotics & Automation Magazine*, 18(1), 24–31.

Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4), 143–166.

Huttenrauch, H. and Eklundh, K.S. (2002). Fetch-and-carry with cero: Observations from a long-term user study with a service robot. In *Proceedings. 11th IEEE International Workshop on Robot and Human Interactive Communication*, 158–163. IEEE.

Kahn, L.E., Averbuch, M., Rymer, W.Z., and Reinkensmeyer, D.J. (2001). Comparison of robot-assisted reaching to free reaching in promoting recovery from chronic stroke. In *Proceedings of the international conference on rehabilitation robotics*, 39–44. IOS Press.

Kanda, T., Ishiguro, H., Imai, M., and Ono, T. (2004). Development and evaluation of interactive humanoid robots. *Proceedings of the IEEE*, 92(11), 1839–1850.

Li, X., Zhang, Y., and Liao, D. (2017). Mining key skeleton poses with latent svm for action recognition. *Applied Computational Intelligence and Soft Computing*, 2017.

Mitra, S. and Acharya, T. (2007). Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(3), 311–324.

Moeslund, T.B., Hilton, A., and Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 104(2-3), 90–126.

Montemerlo, M., Pineau, J., Roy, N., Thrun, S., and Verma, V. (2002). Experiences with a mobile robotic guide for the elderly. *AAAI/IAAI*, 2002, 587–592.

Ng, C.W. and Ranganath, S. (2002). Real-time gesture recognition system and application. *Image and Vision computing*, 20(13-14), 993–1007.

Sempena, S., Maulidevi, N.U., and Aryan, P.R. (2011). Human action recognition using dynamic time warping. In *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*, 1–5. IEEE.

Tapus, A., Maja, M., and Scassellatti, B. (2007). The grand challenges in socially assistive robotics. *IEEE Robotics and Automation Magazine*, 14(1), N–A.

Thobbi, A. and Sheng, W. (2010). Imitation learning of hand gestures and its evaluation for humanoid robots. In *The 2010 IEEE International Conference on Information and Automation*, 60–65. IEEE.

Watanabe, T. and Yachida, M. (1998). Real time gesture recognition using eigenspace from multi-input image sequences. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 428–433. IEEE.

Weingaertner, T., Hassfeld, S., and Dillmann, R. (1997). Human motion analysis: A review. In *Proceedings of the 1997 IEEE Workshop on Motion of Non-Rigid and Articulated Objects (NAM'97)*, 90. IEEE Computer Society.

DOI: 10.17648/sbai-2019-111602