Active Multi-Modal Approach for Enhanced User Recognition in Social Robots

Ninad Kale
Robotics Lab
University at Buffalo
ninadnar@buffalo.edu

Nalini Ratha
Computer Science and Engineering
University at Buffalo
nratha@buffalo.edu

***Abstract*—The field of Human-Robot Interaction (HRI) is swiftly expanding, driven by notable advancements in artificial intelligence (AI). Humanoid robots, now capable of being equipped with advanced AI models, are being considered for a wide array of applications due to their ability to perform complex tasks and interact with humans in both natural and intelligent ways. In the domain of social robotics, the capability to selectively interact with authorized users is crucial for ensuring security and providing personalized user experiences. Unimodal user recognition methods, such as audio-based user recognition, are commonly used in social robots. However, these methods can be susceptible to ambient noise and might exhibit reduced accuracy. Although face recognition modality is often employed to enhance accuracy, the majority of audio-visual person recognition methods are trained on datasets with only a single user in the frame.**

This paper introduces a method for audio-visual user recognition with multiple users in the frame, utilizing an additional sound localization modality. The proposed method is evaluated using a dataset created from interactions between the social robot Pepper and multiple users. The results demonstrated that the proposed method significantly outperformed unimodal user recognition methods.

I. Introduction

In today's ever-evolving landscape of social robotics, the interaction between humans and robots transcends mere functional utility. The burgeoning capabilities of artificial intelligence (AI) have brought about a profound transformation in human-robot interaction, offering unprecedented opportunities for personalized and secure user experiences. At the heart of these interactions lies the promise of tailored, highly responsive engagements with social robots.

Consider robots like Pepper [28], which primarily rely on voice commands to execute tasks and engage with users. However, this particular mode of interaction presents a distinctive challenge, notably in the realm of user identification and authorization. Ensuring the security of such interactions and delivering on the promise of a personalized experience hinges on the robot's ability to discern authorized users from others in the environment. By incorporating effective user identification, these robots can adapt their behavior, facilitating nuanced and individualized interactions, thus pushing the boundaries of human-robot engagement.

Within the domain of voice-interactive robotics, a central concern revolves around the identification and authorization of the user engaging with the robot. While audio embeddings have emerged as a popular solution for user recognition, they come with notable limitations when operating in real-world scenarios. The omnipresent companion of daily life, ambient noise, has the potential to severely disrupt audio signals, ultimately compromising the effectiveness and accuracy of user recognition systems. Similarly face recognition suffers because of lightning, pose and occlusion.

To address these limitations and provide a more resilient solution, it becomes imperative to harness multiple sensory modalities. Many social robots come equipped with not only microphones but also cameras and audio localization capabilities. These features render them prime candidates for the integration of user recognition tasks. Although substantial research exists in this domain, the majority of it falls short in direct applicability to social robots due to the myriad of edge-case scenarios that real-world human-robot interactions encompass.

Moreover, in today's digital age, token-based authentication methods, although useful in various domains, do not seamlessly translate to the realm of social robotics. The need for frictionless and non-invasive authentication mechanisms is paramount, as users increasingly expect a natural interaction with these AI-driven companions. Therefore, this paper aims to address the challenges posed by voice-interactive robots and explore the potential of multi-modal user recognition that circumvents the need for extensive training, aiming for a seamless and adaptive implementation in real-world robot interactions.

The paper is structured as follows: Section II offers an overview of prior research, focusing on uni-modal and multi-modal user recognition in social robotics. In Section III, we dissect the crucial components of face recognition, audio recognition, and audio localization. Our proposed method is presented in the section IV, experimentation and dataset are explained in section V and VI respectively. The results of experiments are presented in Section VII.

II. Related Work

Speaker identification is the task of identifying a speaker from their voice. It is a challenging task, as there is a great deal of variation in human speech, both within and across individuals. In current speaker recognition systems, using a low dimensional fixed-length vector, or speaker embedding, has become the dominant speaker modeling approach.

Over the years, combined with probabilistic linear discriminative analysis (PLDA), i-vector [1] has been the state-of-the-art system for text-independent speaker recognition. Recently, with the development of deep neural networks (DNN), different methods have been explored to increase accuracy. It is demonstrated that a high-performance speaker recognition system can be directly built by training a DNN speaker classifier and extracting embeddings from it. An utterance level DNN speaker embedding, named x-vector [2]–[4] that is produced by a speaker discriminative DNN, has shown better performance than i-vector on a series of speaker recognition tasks.

While in speaker recognition, it is crucial that embeddings from the same identity aggregate and the clusters of different identities are well separated, but the embeddings produced by the DNN are not generalizable enough and performance degradation is observed when evaluated on unseen speakers. Although an entirely end-to-end system can do discriminative embedding learning directly [5]–[8], it requires complicated data preparation such as semi-hard example mining and needs much longer time to train. Embeddings are also susceptible to noisy data which further makes the task of speaker identification more difficult.

Several researchers have shown that the sensor fusion of the audio data and the image, enhances the robustness of speaker recognition [9]–[12]. Literature can be categorized into traditional [13], [14] and deep learning-based approaches [15], [16]. In the work by Das et al. [13], a late fusion strategy is adopted where separate pipelines for audio and visual features are used, and their results are combined. The audio pipeline is based on the x-vector-based speaker embedding, whereas the vision pipeline is based on the ResNet and InsightFace [17].

Deep learning-based approaches report state-of-the-art accuracy [15], [16]. In the work by Vegad et al. [16] two CNN pipelines are used for audio and video-based person recognition. The results obtained from these two pipelines are combined using a weighted average to obtain the final result.

Though the accuracy of speaker identification models is good, using them in real-world scenarios like human-robot interaction is challenging. First, robots will be interacting with multiple humans, with many people talking around them. Most models are trained on clean datasets with a single person in a frame with audio data, so they cannot be directly used in this setup. When multiple people are present in a frame, these modalities are not enough, as we need a way to link audio to one of the people visible in the frame, if it is coming from any of them, or otherwise discard it. Also, many times, some modalities may not be present, maybe the face is occluded, etc. In that case, we need some different modality for person recognition.

While laboratory conditions often yield high recognition rates, real-world applications, especially in social robots, present a plethora of challenges, including varying light conditions, occlusions, and unforeseen noise sources. [21] highlights these challenges and proposes methods to mitigate them. Social robots, like Pepper, introduce unique challenges like dynamic human-robot interactions and varying user distances. Integrating user recognition systems in these robots requires a robust approach that can handle these edge cases.

To address these challenges, we need to develop speaker identification models that are more robust to noise and multiple speakers. We also need to develop methods to correlate audio and visual data to identify the speaker even when some modalities are not present.

Lip sync [18]–[20] is a promising modality for user recognition, as it can be used to identify speakers even in noisy environments. However, lip sync is not always accurate and cannot be used when the speaker's mouth is occluded or when the speaker is not facing directly to the camera. Future work should focus on developing methods to improve the accuracy of lip sync and to address the challenges of occlusion and speaker pose.

Many robots have an audio localization feature, which uses the time difference of arrival of sound waves to estimate the direction from which the sound is coming. This modality is independent of the other two modalities of audio and vision, and it is not susceptible to occlusion or noise. This paper proposes a robust user recognition system that uses audio, visual and sound localization information to identify speakers in a human-robot interface setting, even when multiple users are present in a single frame.

III. Overview

A. Face Recognition

Face recognition techniques can be broadly classified into two main methodologies: classifier-based recognition and embedding-based recognition. In the classifier-based paradigm, the system trains classifiers (e.g., SVMs, Decision Trees, or Neural Networks) directly on facial images. The descriptors or features of these images are often extracted using traditional methods like Haar cascades or HOG (Histogram of Oriented Gradients). However, this method poses a challenge in scalability. Adding a new subject into the system mandates retraining, which can be computationally expensive and time-consuming, especially with increasing dataset sizes. Embedding-based recognition maps facial images into a high-dimensional space using deep learning architectures, producing compact vector representations called "face embeddings". This vector space is carefully crafted such that distances between vectors correlate with facial similarity. The seminal DeepFace paper [22] pioneered this approach using deep convolutional neural networks (CNNs).

Embedding-based recognition is a particularly suitable choice for user recognition in human-robot interaction, as it does not require additional training. Users can conveniently submit facial data during the registration process, which can then be used for authentication, contributing to a seamless and secure interaction experience.

In this paper, we use ArcFace [23] for embedding extraction. ArcFace enhances the discriminative power of embeddings by introducing an additive angular margin penalty to the target logit. This margin penalty ensures that the angular distance between embeddings of the same class is minimized, while that of different classes is maximized. The resultant embeddings, thus, have a smaller intra-class variance and a larger inter-class variance.

B. Audio Recognition

Audio recognition also majorly falls into two categories: feature-based and embedding-based. In the feature-based approach, traditional audio features such as Mel-frequency cepstral coefficients (MFCCs) [24], Chroma features [25] are extracted from the audio signals. These features then serve as input to various classifiers or clustering algorithms to facilitate recognition. While this method has its merits, it can be less adaptive to new or unseen audio variations, making it less versatile in dynamic environments.

The embedding-based paradigm offers a more sophisticated methodology for speaker recognition, employing deep learning models to convert audio signals into high-dimensional vector representations. Much like face embeddings, audio embeddings encapsulate the essence of the audio, providing a compact yet rich representation that can be effectively utilized for recognition tasks.

Our audio embedding extraction hinges on the XVectorSincNet architecture [26], provided through the pyannote library. XVectorSincNet enhances the traditional x-vector TDNN approach by replacing its conventional filter banks with the trainable features of the SincNet architecture [27]. SincNet's pasteurization allow XVectorSincNet to learn adaptive filters directly from audio data, resulting in more discriminative and adaptive audio embeddings.

C. Audio Localisation

To link audio and face embeddings, we can leverage the audio localization modality, which is not susceptible to face occlusion, pose, or ambient noise. Many robots used in human-robot interaction setups have audio localization capability. This feature exploits the time difference of arrival (TDOA) of sound waves at each of the robot's microphones to estimate the azimuthal and elevation angles of the sound source relative to the robot. This information can then be used to transform the audio localization to the camera frame.

The accuracy of the audio localization feature is within 10 degrees according to the dataset. The estimated position in the image frame can be calculated using Gaussian distribution around the mapped pixel from the estimated azimuthal and elevation angle.

Hence, the weight $\mathbf{P}_\text{i}$, which signifies our confidence that the sound is coming from the ith user in the frame, can be given by the following formulation, which is a 2D Gaussian:

$$\mathbf{P}_\text{i} = \mathcal{N}(x_\text{i}, y_\text{i} | x_\text{loc}, y_\text{loc}, \sigma_x, \sigma_y) \tag{1}$$

where x_i, y_i are the mouth coordinates of the ith person in image frame. x_loc and y_loc are the position estimated by audio localisation module. σ_x and σ_y are the standard deviations of the Gaussian distribution.

IV. Proposed Method

In the proposed method, the social robot, Pepper, is used as an experimental tool to collect multimodal data from users. When users pose a question to Pepper, it captures three types of data: audio, visual, and sound localization. It is assumed that the database contains pre-recorded audio and facial data for all participating users.

For audio data, the audio embedding derived from the captured sound is represented as ζ_{record}. To ascertain the identity of a user, the system computes the probability of the recorded audio originating from each user in the database. This is achieved by comparing the ζ_{record} embedding with the stored audio embeddings of each user ζ_{audio_i}. The matching process employs cosine distance to determine the similarity and subsequently calculates the likelihood of a match. Following this computation, we obtain a probability value for each user, indicating the likelihood that the audio recording belongs to them.

$$\mathbf{P}^\text{i}_\text{audio} = \frac{\zeta_{record} \cdot \zeta_{audio_i}}{||\zeta_{record}|| \times ||\zeta_{audio_i}||} \tag{2}$$

Higher the cosine similarity, more is the probability of the audio recording belonging to that user.

Utilizing the sound localization module, we are able to retrieve both the elevation and azimuth angles, which pinpoint the direction from which the sound originated. To integrate this audio information with visual data, these angles are transformed into pixel locations within the image frame. This transformation is tailored to the specific configuration of the robot. The standard deviations, represented as for the horizontal axis σ_x and σ_y for the vertical axis, are computed statistically to provide further insight into the precision of the localization.

For each person detected in the visual frame, the probability that they are the source of the detected sound is computed using Equation (1). The individual with the highest probability is then selected as the likely sound source.

For the selected individual, the probability of the face belonging to each user is determined based on the cosine distance formula, calculated for all users in the database. Let ζ_{loc} be the face embeddings of selected person using audio localisation and ζ_{face_i} be the face embeddings of ith user in database.

$$\mathbf{P}^\text{i}_\text{face} = \frac{\zeta_{loc} \cdot \zeta_{face_i}}{||\zeta_{loc}|| \times ||\zeta_{face_i}||} \tag{3}$$

While embedded fusion using deep learning methods is a common approach to merge audio and visual modalities, it struggles with the absence of one modality. This paper aims to illustrate that a robust system can be developed using simpler fusion strategies.

For each user, a probability is calculated using a weighted average of both the audio and visual probabilities.

$$\mathbf{P}^{\mathrm{i}}_{\mathrm{user}} = \mathbf{w}_{\mathrm{face}} \cdot \mathbf{P}^{\mathrm{i}}_{\mathrm{face}} + \mathbf{w}_{\mathrm{audio}} \cdot \mathbf{P}^{\mathrm{i}}_{\mathrm{audio}} \tag{4}$$

The user corresponding to the maximum probability is identified as the match, provided that this probability exceeds a predefined threshold.

Equal weights yield optimal fusion when both modalities are present; if one is absent, it's excluded by assigning a weight of 1 to the other.

V. Experimentation

A. Pepper Robot

Pepper [28], a humanoid robot from SoftBank Robotics, was used as the base platform for all experiments in this study. Pepper is equipped with audio and visual sensors. The camera image quality is 1280x960 at 5 fps. The Pepper microphones recorded audio of poor quality, which resulted in low accuracy for audio recognition. Therefore, an external microphone was used for audio recording, while the Pepper microphones were only used for sound localization.

Sound localization is performed using four microphones located on the robot's head. The microphones have a sensitivity of 250mV/Pa +/-3dB at 1kHz with a frequency range of 100Hz to 10kHz (-10dB relative to 1kHz). In our testing on a humanoid robot, a sensitivity of 0.9 was found to be optimal for a range of 5 meters. Naoqi library from Softbank Robotics is used for our testing leveraging API-based functions for video feed retrieval and employing interrupt-based functions for efficient sound localization upon sound detection.

VI. Dataset

Since there is no publicly available dataset for audio, video, and sound localization with multiple users in frame for humanoid robots, we collected our own dataset in a laboratory setting.

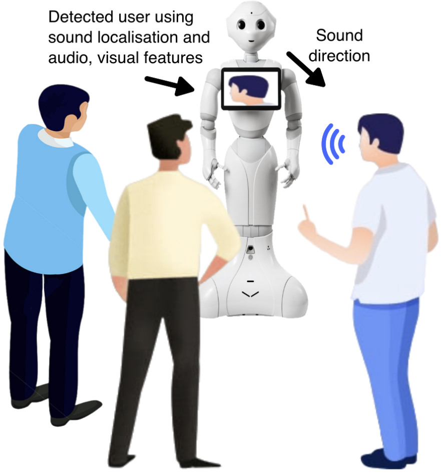
Detected user using sound localisation and audio, visual features
Sound direction

Fig. 1. Pepper, Humanoid Robot interacting with users

We used 25 individuals speaking from multiple locations around Pepper, with multiple users in the frame. The users interacted with Pepper in environments with and without noise, and the dataset was recorded with different Signal-to-Noise Ratio (SNR) levels from -20 to 20 dB.

VII. Results

Table 1 shows the user recognition accuracy utilizing audio-based and multi-modal approaches across varying SNR conditions and user scnarios. In environments where clear audio is prevalent, notably at higher SNR values, user identification accuracy exhibits commendable performance. The value of $\mathbf{P}^{\mathrm{i}}_{\mathrm{audio}}$ predominates the total weightage, emphasizing that audio embeddings significantly enhance user identification, especially in acoustically favorable conditions.

As noise intrudes and SNR diminishes, the unimodal approach experiences a decline in accuracy. Contrastingly, our proposed method demonstrates steadfast user recognition accuracy, even amidst elevated SNR, underlining its consistency and potential applicability in noise-afflicted, real-world scenarios.

TABLE I
User Recognition Accuracy with Respect to SNR

SNR	Audio based (%)	Multi Modal single user	Multi Modal multiple users
-20dB	83%	95%	91.3%
-10dB	91%	97%	95.6%
0	97.2%	98.8%	98.1%
10dB	99.65%	100%	99.4%
20dB	99.8%	100%	99.4%

Table 2 elucidates the relationship between user localization accuracy and various SNR levels. A discernible trend emerges where an increase in ambient noise, and consequently a decrease in SNR, correlates with a reduction in localization accuracy.

TABLE II
User Localisation Accuracy at Different SNR Levels

SNR	User Localisation Accuracy (deg)	Confidence Score
-20dB	13.3	0.7
-10dB	12.8	0.75
0	10.9	0.89
10dB	7.3	0.95
20dB	7.2	1

Optimized results have been attained through careful adjustment of audio sensitivity parameters, demonstrating particular effectiveness for the Pepper humanoid robot in sustaining accurate user localization, even in environments with acoustical challenges.

VIII. Conclusion

Consequently, it can be deduced from the results that the proposed method exhibits resilience to elevated noise levels and possesses the capability to scale effectively with multiple users in the environment. This renders it a pragmatic approach for implementation in real-world scenarios, especially for social robots engaging with groups of individuals.

References

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788–798, 2011.

[2] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, “Deep neural network-based speaker embeddings for end-to-end speaker verification,” in 2016 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2016, pp. 165–170

[3] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep neural network embeddings for text-independent speaker verification,” in Proc. Interspeech 2017, 2017, pp. 999–1003

[4] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 5329–5333.

[5] G . Heigold, I. Moreno, S. Bengio, and N. Shazeer, “End-toend text-dependent speaker verification,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 5115–5119.

[6] C. Zhang and K. Koishida, “End-to-end text-independent speaker verification with triplet loss on short utterances.” in Interspeech, 2017, pp. 1487–1491

[7] S.-X. Zhang, Z. Chen, Y. Zhao, J. Li, and Y. Gong, “End-to-end attention based text-dependent speaker verification,” in 2016 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2016, pp. 171–178

[8] Z. Huang, S. Wang, and Y. Qian, “Joint i-vector with end-to-end system for short duration text-independent speaker verification,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 4869–4873

[9] Wen, Y., Ismail, M.A., Liu, W., Raj, B., Singh, R.: Disjoint mapping network for cross-modal matching of voices and faces. In: Proceedings of the International Conference on Learning Representations (ICLR), pp. 1–17, May 2019

[10] Nawaz, S., Janjua, M.K., Gallo, I., Mahmood, A., Calefati, A.: Deep latent space learning for cross-modal mapping of audio and visual signals. In: Proceedings of the Digital Image Computing: Techniques and Applications (DICTA), pp. 1–7, December 2019

[11] Sell, G., Duh, K., Snyder, D., Etter, D., Garcia-Romero, D.: Audio-visual person recognition in multimedia data from the Iarpa Janus program. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3031–3035, April 2018

[12] Tao, R., Das, R.K., Li, H.: Audio-visual speaker recognition with a cross-modal discriminative network. In: Proceedings of Annual Conference of the International Speech Communication Association, (INTERSPEECH), pp. 2242–2246, October 2020

[13] Das, R.K., Tao, R., Yang, J., Rao, W., Yu, C., Li, H.: HLT-NUS submission for 2019 NIST multimedia speaker recognition evaluation. In: Proceedings of the APSIPA, Annual Summit and Conference, pp. 605–609, December 2020

[14] Sadjadi, S., Greenberg, C., Singer, E., Olson, D., Mason, L., Hernandez-Cordero, J.: The 2019 NIST audio-visual speaker recognition evaluation. In: Proceedings of the Speaker and Language Recognition Workshop: Odyssey 2020, pp. 266–272 (2020)

[15] Geng, J., Liu, X., Cheung, Y.M.: Audio-visual speaker recognition via multi-modal correlated neural networks. In: Proceedings of 2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW), pp. 123–128, October 2016

[16] Vegad, S., Patel, H.P.R., Zhuang, H., Naik, M.R.: Audio-visual person recognition using deep convolutional neural networks. J. Biometrics Biostatistics 8, 1–7 (2017)

[17] Guo, J., Deng, J., Lattas, A., Zafeiriou, S.: Sample and computation redistribution for efficient face detection. arXiv preprint arXiv:2105.04714, May 2021

[18] J. S. Chung, A. Senior, O. Vinyals and A. Zisserman, ”Lip Reading Sentences in the Wild,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 3444-3453, doi: 10.1109/CVPR.2017.367.

[19] K. Vayadande, T. Adsare, N. Agrawal, T. Dharmik, A. Patil and S. Zod, ”LipReadNet: A Deep Learning Approach to Lip Reading,” 2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC), Dharwad, India, 2023, pp. 1-6, doi: 10.1109/ICAISC58445.2023.10200426.

[20] A. Adeel, M. Gogate, A. Hussain and W. M. Whitmer, ”Lip-Reading Driven Deep Learning Approach for Speech Enhancement,” in IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 5, no. 3, pp. 481-490, June 2021, doi: 10.1109/TETCI.2019.2917039.

[21] Rafizah Mohd Hanifa, Khalid Isa, Shamsul Mohamad, A review on speaker recognition: Technology and challenges, Computers & Electrical Engineering, Volume 90, 2021, 107005, ISSN 0045-7906, https://doi.org/10.1016/j.compeleceng.2021.107005.

[22] Y. Taigman, M. Yang, M. Ranzato and L. Wolf, ”DeepFace: Closing the Gap to Human-Level Performance in Face Verification,” 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 1701-1708, doi: 10.1109/CVPR.2014.220.

[23] Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4690-4699).

[24] Z. K. Abdul and A. K. Al-Talabani, ”Mel Frequency Cepstral Coefficient and its Applications: A Review,” in IEEE Access, vol. 10, pp. 122136-122158, 2022, doi: 10.1109/ACCESS.2022.3223444.

[25] Abraham, J.V.T., Khan, A.N. & Shahina, A. A deep learning approach for robust speaker identification using chroma energy normalized statistics and mel frequency cepstral coefficients. Int J Speech Technol (2021). https://doi.org/10.1007/s10772-021-09888-y

[26] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, ”X-Vectors: Robust DNN Embeddings for Speaker Recognition,” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 5329-5333, doi: 10.1109/ICASSP.2018.8461375.

[27] M. Ravanelli and Y. Bengio, ”Speaker Recognition from Raw Waveform with SincNet,” 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 2018, pp. 1021-1028, doi: 10.1109/SLT.2018.8639585.

[28] https://www.aldebaran.com/en/pepper