Multimodal Biometric Identification System for Mobile Robots Combining Human Metrology to Face Recognition and Speaker Identification

Simon Ouellet, François Grondin, Francis Leconte and François Michaud

Abstract— **Recognizing a person from a distance is important to establish meaningful social interaction and to provide additional cues regarding the situations experienced by a robot. To do so, face recognition and speaker identification are biometrics commonly used, with identification performance that are influenced by the distance between the person and the robot. This paper presents a system that combines these biometrics with human metrology (HM) to increase identification performance and range. HM measures are derived from 2D silhouettes extracted online using a dynamic background subtraction approach, processing in parallel 45 front features and 24 side features in 400 ms compared to 38 front and 22 side features extracted in sequence in 30 sec by using the approach presented by Lin and Wang [1]. By having each modality identify a set of up to five possible candidates, results suggest that combining modalities provide better performance compared to what each individual modality provides, from a wider range of distances.**

I. INTRODUCTION

Being able to identify people is a key element in creating personalized interactions [2], and the same certainly holds for human-robot interaction (HRI). People identification is essential to greet a person, to store information regarding an individual or to direct conversation towards topics of interest. Identification of humans by measurable characteristics is referred to as biometrics. Hard biometric identifiers are associated to permanent and distinct traits unique to each individual, such as face, voice, fingerprint, iris, DNA, etc. [3], while soft biometric traits are not (e.g., colour of eye, hair, beard or skin, height and weight, body shape, clothes) [4], [5] and require less computation power to process [4].

In HRI, face recognition is one of the most common hard biometrics used [6], [7], [8], [9]. Unfortunately, face recognition approaches are sensitive to illumination conditions and the quality (size, orientation, segmentation) of the region of interest to analyze [10]. Speaker identification is reliable in quiet environments or when using a microphone in close proximity of the person, conditions that are uncommon in natural settings [11]. Other hard biometric identifiers such as fingerprint, iris and DNA are accurate but cannot be used from a distance. As a solution, combining face recognition and speaker identification with other vision-based soft biometrics can help provide more robustness in human identification at a distance and in unconstrained and dynamic environments [12]. For instance, Lawson and Martinson [7] uses a Markov Logic Network to fuse soft biometric indicators (nose, forehead and clothing images taken relative to the detected face position) with face recognition and speaker identification, demonstrating 74% recall and 97% precision, with people facing forward 1 m to 2 m away from the robot. However, pose requirements make interaction restricted to a limited and constrained space nearby the robot.

To increase the interaction space, this paper presents a system that combines face recognition and speaker identification with human metrology (HM) features (e.g., body shape, anthropometric measurements and geometrical features generated from these measurements). According to Adjeroh and al. [13], human metrology can be defined using 10 measures of lengths (arm, head), breadths (head, shoulder), circumferences (armscye, chest, neck base, waist), stature and weight. Circumference measures are difficult to extract from 2D images and require precise 3D modelling [13]. However, HM measures such as height [14] and silhouette [15] can be perceived from 2D images. Lin and Wang [1] developed an automated body feature extraction method from 2D images, generating 38 front and 22 side features from a silhouette in approximately 30 seconds. They used a clockwise sequential extraction process initiated from a starting point (the top of the head) to then follow the silhouette and extract the desired features in a specific order.

HM has not yet been used in HRI for people identification, and could be quite beneficial if processing time could be decreased substantially. The main contribution of this paper is to present improvements made to Lin and Wang's approach, such as using additional features, following an anthropometry-based features search and exploiting a parallel extraction process to derive measures independently from one another, making it fast enough to be used for online processing and minimizing latency. In addition, a dynamic background substraction was used in order to enable HM extraction in unconstrained environment. To illustrate potential benefits of the approach, fusion of face recognition, speaker identification and HM for trials conducted in controlled conditions is presented, simply by weighting the respective identification results according to the distance between the person and the robot.

This work was supported by the Natural Sciences and Engineering Research council of Canada (NSERC) and the Fonds de recherche du Québec – Nature et technologies (FRQNT).

S. Ouellet, F. Grondin, F. Leconte and F. Michaud are with the Interdisciplinary Institute of Technological Innovation (3IT) and the Department of Electrical Engineering and Computer Engineering, Université de Sherbrooke, 2500 boul. Université, Sherbrooke, Québec, CANADA {Simon.Ouellet, Francois.Grondin2, Francis.Leconte, Francois.Michaud}@usherbrooke.ca

The paper is organized as follows. Section II presents our system and the biometric modules. Section III describes the experimental setup and the system's configuration for the trials. Section IV provides identification results in relation to the distance between the robot and the person, for each biometric module and for their combined results.

II. Online Human Identification System

Figure 1 illustrates our system's overall architecture. It consists of four primary modules: one for each biometric modality used (Face recognition, Speaker identification, and HM), and a Fusion module. Biometric modules basically examine features from different perceptual modalities to find matches with training data stored in the databases. Each module i gives an ordered list of identification candidates k from a set of K models, along with their confidence level $s_i(k)$. Fusion consists of combining the confidence levels $s_i(k)$ from the biometric modules to provide the overall confidence level $s(k)$ for each possible identity K.

A. Face Recognition

Face recognition is done using the FaceRecognizer class from OpenCV 2.4 and a Kinect infrared depth-sensing camera (with a range of 0.8 m to 4 m). The Haar Cascade Face Detector, also known as the Viola-Jones method [16], is used for face detection. Once a face has been detected, a histogram equalization is applied to standardize the brightness and contrast of the image. Then, the Eigenfaces method (also known as Principal Component Analysis (PCA)), is applied on the pre-processed facial image [17]. The number of extracted eigenfaces is based on the number of faces in the database. A nearest neighbour method is used to identify the most likely candidate(s): confidence levels $s_0(k)$ are evaluated according to (1) using the Euclidean distance δ between the perceived and the database features, where N is the number of trained faces.

$$s_0(k) = 1 - \frac{1}{255}\sqrt{\frac{\delta}{N}} \tag{1}$$

B. Speaker Identification

Speaker identification is performed using ManyEars [18] and WISS [19]. Raw audio data from a microphone array are sent to ManyEars, a source localization, tracking and separation system designed for mobile robots [18]. The separated and postfiltered audio streams are then fed to WISS, a speaker identification system for mobile robots operating in noisy and reverberant environments [19]. WISS generates a set of speech features and also estimates the additive and convolutive noises in the room. A Parallel Model Combination (PMC) technique is proposed to update each speaker model initially trained in a clean environment, such that these models match the actual environment. Each model is represented by a set of clusters, which are moved according to additive and convolutive noises. The deviation between features and the clusters of each model k is represented by the variable $h(k)$. The speech features are compared to the updated models and a confidence level $s_1(k)$ is provided for each model k, as expressed by (2). The indexes k_1 and k_2 correspond to the models with the smallest ($h(k_1)$) and second smallest ($h(k_2)$) deviations. The parameters α_1 and β_1 are set to 0.02 and 0.01, respectively. The weighted rate of good speaker identifications is 96% for a signal-to-noise ratio (SNR) of 16 dB and 84% at a SNR of 2 dB when pink noise is observed [19].

$$s_1(k) = \begin{cases} \left(1 + \exp\left[-\frac{(h(k_1)-h(k_2))-\alpha_1}{\beta_1}\right]\right)^{-1} & k = k_1 \\ 1 - s_1(k_1) & k = k_2 \\ 0 & k \neq k_1, k_2 \end{cases} \tag{2}$$

C. HM

Human metrology extraction starts by subtracting the image background to obtain the foreground silhouette, as illustrated by Fig. 2. Failure to subtract the background precisely impacts the precision of the foreground silhouette, and therefore influences HM identification performance. Using point cloud data from the Kinect would help extracting the foreground and pose, but would limit the use of HM to less than 4 m. Two methods for background segmentation are used: one to generate the training data set used for matching, and one used to acquire image in real-world settings. As done by Lin and Wang [1], the training data set is created from images (different from the ones used for face recognition) of people wearing white underware in front of a black background and under standard illumination conditions, to get precise HM measures (which would not be possible if casual clothing was allowed). Front and side images are taken with people adopting a standard posture with their limbs straight and arms apart from the torso, to obtain a precise silhouette and estimate all body features for the training process. A basic binary thresholding is used to classify pixels that belong to the object from those of the background. In real-world use, we chose to use ViBe, an open source universal background subtraction algorithm [20], for background segmentation (while Lin and Wang approach always assumes that people are taking specific poses in front of a black background). ViBe takes a pre-processed image as input and outputs a binary image of the estimated foreground. It works with possible occlusions with obstacles and in front of different types of static or changing backgrounds. It enables the system to detect motion when the robot is immobile, and can easily reinitialize the background model once the robot starts to move again. Taking a pose as for the training data is not required because the system attempts to identify people even if a subset of body features are visible; it does however help identify the largest possible set of features.

Then, also as in [1] and as shown in images 2e to 2h, a Canny edge detector [21] is used to do foreground boundary extraction for silhouette segmentation. Internal holes in the foreground silhouette are filled so that edge pixels are linked to a continuous and closed silhouette curve. Each pixel position in the silhouette is transformed using Freeman's 8-connected chain codes to obtain the direction for the current

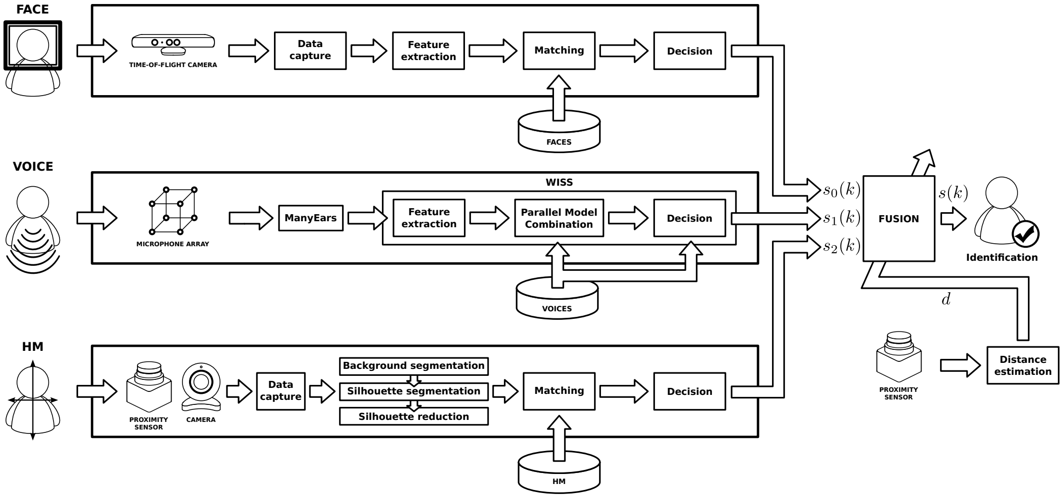
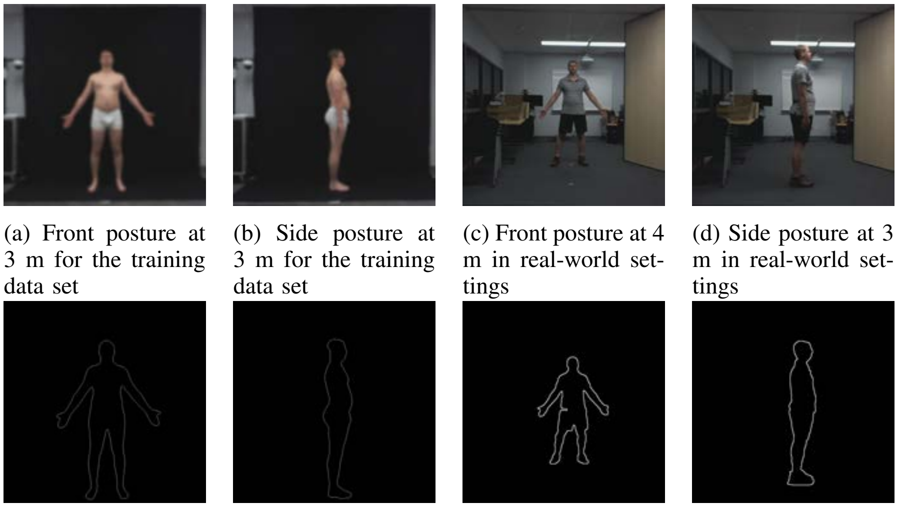


Fig. 1: System architecture



(a) Front posture at 3 m for the training data set

(b) Side posture at 3 m for the training data set

(c) Front posture at 4 m in real-world settings

(d) Side posture at 3 m in real-world settings

(e) Segmented silhouette from front posture for the training data set

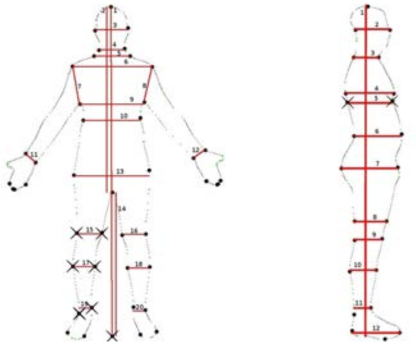
(f) Segmented silhouette from side posture for the training data set

(g) Segmented silhouette from front posture in real-world settings

(h) Segmented silhouette from side posture in real-world settings

Fig. 2: Background segmentation and silhouette extraction

pixel [22]. Finally, to reduce the computational load and memory usage for feature extraction, a silhouette reduction method minimizes the chain code length by detecting straight lines and removing duplicated directions, keeping only the points where the direction changes in relation to the starting point. Once the silhouette is reduced, all silhouette features are extracted using the 90° directional change rule from Lin and Wang [1], reducing the number of silhouette contour points by approximately 90%. But contrarily to Lin and Wang approach, HM is done over 45 front features and 24 side features (compared to 38 front and 22 side features for [1]) shown in Fig. 3a and Fig. 3b, respectively, with each feature represented by a dot and new features by a X on the silhouettes. Feature estimation from the segmented silhouettes follows an anthropometric-based method developed to evaluate the optimal position of each measure to be used for matching. The optimal location of each feature in the image is compared with the average white U.S. american anthropometric measures [23] using a nearest neighboor method. If the distance between the anthropometry position being evaluated and the estimated feature is higher than a threshold of l_f pixels, the feature is rejected and not used for identification. The threshold l_f is set based on the standard deviation between each optimal feature location and each detected features f from all training data. A parallel feature extraction process is used to independently detect feature without relying on the previous detected feature. So instead of having to stop the feature extraction process as soon as one one feature is undetectable or if the deviation error from the anthropometric optimal position is too large (as in Lin and Wang), the parallel feature extraction process makes the approach robust to missing feature by detecting all possible



(a) Front measures (b) Side measures

Fig. 3: HM Measurements

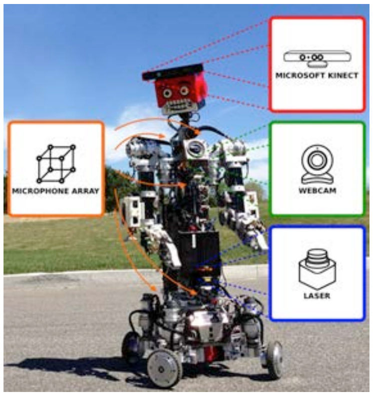


Fig. 4: IRL-1 experimental platform.

features. It also decreases processing time from 30 seconds to 400 ms for a given input image, processed in a specific thread initiated when the image is acquired.

From these detected features, a total of 20 front measures and 12 side measures are extracted in the process, shown in Fig. 3a and Fig. 3b. Only measures with the available features are evaluated. These measures are derived from the x axis distance for horizontal measures, the y axis distance for vertical measures, and the euclidean distance for straight measures. A conversion table is created by taking images, from 40 cm to 6 m with 20 cm increment, of a chessboard and calculating the size of a line at each step, and then applying a linear interpolation, as the real length of a line is known. Laser range finder data are used to improve the precision at close range (between 0.8 m and 3.5 m, with a Hokuyo UTM-04LX laser range finder) using the distances of both legs and of the torso projected axis at the centre of both legs on the ground.

Matching is then conducted using the following global measure deviation method. Each estimated measure $m_e(p)$ is compared to the modelled measures $m_k(p)$ ($p = 0, 1, \ldots, P-1$) from the training data to evaluate the summed standard deviation $g(k)$ over the K models in the training set. Once all training models are evaluated, an ordered list of potential candidates C is selected from the list of models with the lowest global feature deviations. The global feature deviation of the $(C+1)$ best model (the first ordered candidate outside the list) is used as a reference g_{ref}. A sigmoid function expressed by (3) is then used to generate $s_2(k)$, with $\alpha_2 = 10$ and $\beta_2 = 5$, to generate the confidence score for each C candidate. Front and side measures are matched independently and a confidence value is calculated for each posture depending on the current pose of the person.

$$s_2(k) = \left(1 + \exp\left[-\left\{(g(k) - g_{ref})^2 - \alpha_2\right\}/\beta_2\right]\right)^{-1} \quad (3)$$

D. Fusion

The objective of this module is to combine the identification results of the three biometric modalities, taking into consideration that each one is independent, asynchronous and can provide valid identifications in complementary conditions. For instance, for face recognition the person must be close to the robot, while speaker identification or HM may be reliable at larger distances. The Fusion module receives an ordered list of the best potential identification candidates from each biometric modalities, expressed by $s_i(k)$. Each modality list of identified candidates with their confidence levels $s_i(k)$ is normalized using the $\tanh$ normalization estimator (also known as the Hampel estimator [24]) according to [25]. Hampel estimators reduce the influence of the points at the tails of the distribution during the estimation. Equation (4) evaluates $s(k)$ as a weighted sum on the confidence levels $s_i(k)$. Weights $w_i(d)$ are chosen according to the perceived distance d between the person and the robot, and their sum over i is one. The variable $c_i(k)$ represents the rank in the ordered list of $s_i(k)$ (e.g., $c_i(k) = 0$ if the model k has the highest confidence level $s_i(k)$). It is used to decrease the influences of candidates with low confidence levels. The final identification provided by the Fusion module corresponds to k with the highest $s(k)$ and with $s(k) >= T$, with T being a threshold set to designate a valid identification.

$$s(k) = \sum_{i=0}^{2} s_i(k) w_i(d) \quad (4)$$

III. Experimental Settings

Figure 4 illustrates the platform IRL-1 used for the trials. IRL-1 use a Kinect camera to detect the presence of a person using a RGB 640 × 480 image at 30 Hz for face recognition. A microphone array with eight channels is used to generate an audio stream for the active speaker. The multi-channel signals are sampled with the 8SoundUSB sound card at 48000 samples per second [26]. A high definition camera (Logitech c910) is used to record images for HM with a 1280 × 960 resolution at 10 Hz, which greatly reduces the level of distortion. These images are encoded as a 8-bit per channel BGR format. A pinhole model is also used to cancel lens distortions. The digital camera is fixed to the robot body, 1 m above the ground. The IMU (Inertial Measurement Unit) on IRL-1 is used to stop image acquisition when slight oscillations are observed. A Hokuyo UTM-30LX laser range finder is used to measure the distance with the person.

Training data were acquired for 22 male participants. Data acquisition for face recognition and speaker identification were done with each participant sitting 1.5 m away from the recording devices. For face recognition training data, images were captured for 30 seconds, in order to obtain 150 images, with the participant doing the following: 1) Look directly to the camera for the first 5 seconds, adopting a neutral pose with no facial expressions; 2) Look in each direction (left, right, top, bottom) at a 45° angle for at most 2 to 3 sec each; 3) Look directly to the camera while talking and changing facial expressions for the last 13 to 17 seconds. During this training, one model is created for each participant, creating a list 150 of eigenfaces for each images acquired. For speaker identification training data, each participant had to read out loud random passages from a book, for 2 min. For HM training data, participants stood 3 m away from the camera, taking front and side poses for 30 sec each. One model per participant is generated for each pose.

Trials in real-world conditions were done with 7 of these participants, selected randomly and wearing casual clothing. For face recognition, N was set to 3300.The reverberant time decay of the room (RT60) was 300 msec. The average SNR for each microphone was estimated at approximately 3 dB due to the wideband noise of the fans on the robot. Each trial started with the participant being out of IRL-1's sight, then moved to the 1 m mark on the floor in front of the robot, keeping each pose (front or side) twice for 30 seconds, while talking and facing the Kinect when adopting the front pose. This allowed us to characterize the performance of the system in the best possible and in controlled conditions. IRL-1 remained immobile to facilitate background segmentation. T was set empirically to 87.5%.

IV. RESULTS

Performance can be expressed in terms of recall, precision and accuracy, as defined in (5), with t_p the number of valid acceptation (true positives), t_n the number of failure to reject invalid result (true negative), f_p the number of failure to accept valid result (false positives), and f_n the number of failure to reject false result (false negatives). Weight optimization was done by evaluating performance for all weight combination (0.01 increment), and by selecting the ones minimizing the false acceptance rate (FAR) and the false reject rate (FRR) while taking in account the failure to capture rate (FTC) (or recall). Table I presents the weights derived. Performance are evaluated using the recognition data of the 7 participants for all distances, both poses and both sequences. The system provides a list of all trained candidates, ordered according to their confidence level based on FAR. Results are presented in relation to distances for each biometric modality and for their combined results.

$$Recall = \frac{t_p}{t_p + f_n}, Precision = \frac{t_p}{t_p + f_p}, \\ Accuracy = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \tag{5}$$

TABLE I: $w_i(d)$ for each modality

Distance	Weights		
(m)	Face	Speaker	HM
1	0.9	0.1	0
2	0.684	0.316	0
3	0	0.582	0.418
4	0	0.421	0.579
5	0	0.667	0.333
6	0	0.785	0.215

Table II gives Recall, Precision and Accuracy performance for face recognition, speaker identification and HM. For face recognition, recall goes from 79% at 1 m to 22% at 2 m. Good accuracy is observed at 1 m but quickly drops at 2 m, meaning that more candidates are required to compensate for the difficulty of recognizing faces (which become smaller as people are farther away from IRL-1). Farther to 2 m, no faces (e.g., Recall = 0) can be detected in any images. As for precision, if a face in detected then it can be identify without the need of segmenting or extracting information in order to identify the individual, thus the 100% precision. For speaker identification, ManyEars detects a sound source and tracks it until the source becomes silent for many seconds. All the frames in the audio stream generated are used for speaker identification, and WISS deals internally with time-frequency masks for speech activity. This led to a recall and precision rate of 100% for all trials. Performance for this modality are affected by audible noises made by the onboard fans located on IRL-1's torso and the level of reverberation in the room. As the distance between IRL-1 and the participant increased, accuracy decreased because the amplitude of the audible speech decreased. Regarding HM, recall can be greater than 40% at distances of 2 to 6 m, from which images of the entire body can be processed, depending on the individual height. The best precision and accuracy with the highest recall are observed at 3 m where each feature are further apart and the effect of features deviation from the optimal position have less impact than at higher distance. At 1 m and 2 m, some features are not visible and some overlap with other features, decreasing precision. For larger distances, accuracy decreased because the perceived silhouettes were smaller, making it harder to identify a person from its silhouette. Moreover, casual clothing in real life makes identification less accurate.

TABLE III: Fusion Identification Performance

Distance	Recall/Precision	Accuracy (k) (%)				
(m)	(% / %)	1	2	3	4	5
1	91 / 100	93	96	100	100	100
2	58 / 100	51	72	90	93	94
3	78 / 99	43	79	79	86	86
4	71 / 99	29	50	64	64	71
5	74 / 97	14	43	50	57	64
6	79 / 91	0	21	21	36	64

TABLE II: Modality Recognition Performance

	Face						Speaker						HM					
Dist.	Recall / Prec.	Accuracy (k) (%)					Recall / Prec.	Accuracy (k) (%)					Recall / Prec.	Accuracy (k) (%)				
(m)	(% / %)	1	2	3	4	5	(% / %)	1	2	3	4	5	(% / %)	1	2	3	4	5
1	79 / 100	96	96	96	97	97	100 / 100	43	64	71	71	71	0 / 0	0	0	0	0	0
2	100 / 22	37	42	54	81	89	100 / 100	29	43	43	50	50	24 / 47	16	31	31	31	31
3	0 / 0	0	0	0	0	0	100 / 100	14	50	50	57	57	47 / 99	37	44	49	51	53
4	0 / 0	0	0	0	0	0	100 / 100	14	36	36	36	43	42 / 98	26	34	36	47	54
5	0 / 0	0	0	0	0	0	100 / 100	7	36	36	43	50	44 / 96	7	16	19	27	33
6	0 / 0	0	0	0	0	0	100 / 100	0	14	14	29	50	47 / 83	5	7	8	11	17

Table III presents the overall fusion identification performance. An accuracy of 93% is achieved for a single person at 1 m. Accuracy goes down to 0% at 6 m because the identification confidence is not high enough to pass the threshold for any person. Accuracy reaches 100% at 1 m with $k = 3$, and a rate of 64% is achieved at a distance of 6 m with $k = 5$. WISS produces an identification list every second, while the face and HM produce an identification list for every frame in which the modality is detected. Thus, the overall system identifies individual every 200 ms with all combined data acquired.

V. CONCLUSION AND FUTURE WORK

This paper presents a multimodal biometric identification system that uses human metrology measures to complement face recognition and speaker identification modalities. Results suggest that the combination of these biometrics modalities can improve precision over a larger range of distances between the person and the robot, compared to each modality taken individually. Our intent with this work was to demonstrate the potential benefits of using HM as a modality for online biometric identification. In future work, we intend to study how to optimize the weighted influences of each biometric modality by using for instance Markov Logic Network [7], extend the work to create a 3D human body model in virtual space, and to validate the use of the approach in natural settings, indoor and outdoor, with the participants and the robot moving.

ACKNOWLEDGMENT

This works has been approved by Ethics Committee for Research Education and Social Sciences, Univ. Sherbrooke.

REFERENCES

[1] Y.-L. Lin and M.-J. J. Wang, “Automated body feature extraction from 2D images,” *Expert Systems with Applications*, 38(3): 2585–91, 2011.

[2] M. Fishbein and I. Ajzen, *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research.* Addison-Wesley, 1975.

[3] L. Wang, “Some issues of biometrics: Technology intelligence, progress and challenges,” *International Journal of Information Technology and Management*, 11(1/2): 72 – 82, 2012.

[4] A. K. Jain, S. C. Dass, and K. Nandakumar, “Can soft biometric traits assist user recognition?” in *Biometric Technology for Human Identification*, vol. 5404. SPIE, 2004, pp. 561–72.

[5] A. Dantcheva, C. Velardo, A. D’Angelo, and J. L. Dugelay, “Bag of soft biometrics for person identification. new trends and challenges,” *Multimedia Tools and Applications*, 51: 739–777, 2011.

[6] R. Stiefelhagen, H. K. Ekenel, C. Fugen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit, and A. Waibel, “Enabling multimodal human-robot interaction for the Karlsruhe humanoid robot,” *IEEE Trans. on Robotics*, 23(5): 840–51, 2007.

[7] W. Lawson and E. Martinson, “Multimodal identification using Markov logic networks,” in *Proc. 9th IEEE Int. Conf. on Automatic Face and Gesture Recognition*. 2011, pp. 65–70.

[8] K. Mykoniatis, A. Angelopoulou, K. Schaefer, and P. Hancock, “Cerberus: The development of an intelligent autonomous face recognizing robot,” in *Proc. IEEE Int. Systems Conf.*, vol. 1, 2013, pp. 376–80.

[9] Y. Zhang, K. Hornfeck, and K. Lee, “Adaptive face recognition for low-cost, embedded human-robot interaction,” in *Proc. IAS Int. Conf. on Intelligent Autonomous Systems*, vol. 5404, 2013, pp. 863–72.

[10] A. Jain, L. Hong, and S. Pankanti, “Biometric identification,” *Communications of the ACM*, 43(2): 90–98, 2000.

[11] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, “Robust speaker recognition in noisy conditions,” *IEEE Trans. on Audio, Speech, and Language Processing*, 15(5):1711–23, 2007.

[12] D. A. Reid and M. S. Nixon, “Imputing human descriptions in semantic biometrics,” in *Proc. ACM Workshop on Multimedia in Forensics, Security and Intelligence*, vol. 1, 2010, pp. 25–30.

[13] D. Adjeroh, D. Cao, M. Piccirilli, and A. Ross, “Predictability and correlation in human metrology,” in *Proc. IEEE Int. Workshop on Information Forensics and Security*, 2010, pp. 1–6.

[14] C. Madden and M. Piccardi, “Height measurement as a session-based biometric for people matching across disjoint camera views,” in *Proc. Image and Vision Computing New Zealand*, 2005, p. 29.

[15] N. Mallikarjun Reddy Burri, *Exploring the Use of Human Metrology for Biometric Recogition.* ProQuest, 2007.

[16] P. Viola and M. Jones, “Robust real-time object detection,” in *Proc. International Journal of Computer Vision*, 2001, pp. 137–154.

[17] M. Turk and A. Pentland, “Face recognition using eigenfaces,” in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, 1991, pp. 586–591.

[18] F. Grondin, D. Létourneau, F. Ferland, V. Rousseau, and F. Michaud, “The ManyEars open framework,” *Autonom. Robots*, pp. 1–16, 2013.

[19] F. Grondin and F. Michaud, “WISS, a speaker identification system for mobile robots,” in *Proc. IEEE Int. Conf. on Robotics and Automation*, 2012, pp. 1817–1822.

[20] M. V. Droogenbroeck and O. Paquot, “Background subtraction: Experiments and improvements for ViBe,” in *Proc. Computer Vision and Pattern Recognition Workshops*, 2012, pp. 32–37.

[21] J. Canny, “A computational approach to edge detection,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, 8(6): 679–698, 1986.

[22] H. Freeman and L. S. Davis, “A corner-finding algorithm for chain-coded curves,” *IEEE Trans. on Computers*, C-26(3): 297–303, 1977.

[23] R. Contini, “Body segment parameters, Part II,” *Artificial Limbs*, vol. 16, no. 1, pp. 1–19, 1972.

[24] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*. Wiley, 1986.

[25] A. R. Anil K. Jain, Karthik Nandakumar, “Score normalization in multimodal biometric systems,” *Journal of Pattern Recognition*, 38(12):2270—2285, 2005.

[26] D. Abran-Côté, M. Bandou, A. Béland, G. Cayer, S. Choquette, F. Gosselin, F. Robitaille, D. Telly Kizito, F. Grondin, and D. Létourneau. (2012) USB Synchronous Multichannel Audio Acquisition System. [Online]. Available: http://sourceforge.net/projects/eightsoundsusb/files/TechnicalPaper/