Online Audiovisual Signature Training for Person Re-identification

François-Xavier Decroix[13] Frédéric Lerasle[23] Julien Pinquier[1] Isabelle Ferrané[1]

[1] Université de Toulouse, UPS, IRIT {decroix, pinquier, ferrane}@irit.fr
[2] CNRS, LAAS, lerasle@laas.fr [3] Université de Toulouse, UPS, LAAS

Abstract

In intelligent environments, activity detection is a necessary pre-processing step for adaptive energy management and interaction with humans. To characterize the interactions between individuals or between an individual and the infrastructure of a building, a re-identification process is required and using multimodal models improves its robustness. In this paper, we propose a method for audiovisual fusion, which introduces a novel confidence index of audio-video salience zones, for training an audiovisual signature of a person within a sparse network of cameras and microphones.

1 Introduction

Person re-identification is defined as the process of matching new observations of a person detected by a sensor or a network of sensors with previous observations of the same person. Unlike person identification task, no prior information is required and this task can be proceeded online, since no label are intended to be put on the detections. To discriminate a person among others, a signature is required.

With the flourishing expansion of ambient sensors in professional as well as in private lives, it can serve many applications in surveillance [15], multimedia information retrieval and activity detection. To benefit from the complementarity of the sensors percepts, multimodal approaches combine RGB images with depth and thermal features [17] or with RFID data [10], improving robustness of a person signature, since visual appearance is not always discriminative. Nevertheless mixing audio-based and video-based information remains challenging due of the dissemblance of these two modalities. Mimicking human activity and interaction, which are inherently multimodal, audiovisual fusion has been opening its area of applications from human computer interface to intelligent vehicles [26] or smart homes [4].

In this paper, we present a new strategy of audiovisual fusion to build a bimodal signature of a person in a context of indoor activity detection in small rooms where installed sensors have partially joined fields to limit the instrumentation of the place. A novel confidence index in joint audiovisual domain is described to build an audiovisual signature of a person, and is validated on our own audiovisual database.

The remainder of the paper is organized as follows. Both audio and visual signatures are discussed in section 2. Our approach for audiovisual fusion of the signatures is described in section 3. Experimental results and conclusion are given in sections 4 and 5 respectively.

2 Audio and Visual Signatures

The whole architecture of our system is shown in Figure 1. The sub-synoptic in blue is explained in this section. The two other parts, circled in green and red respectively, are detailled in section 3.

Visual appearance and voice tone are two uncorrelated modalities: one cannot be predicted by observing the other. Information of gender or age is exploitable in particular cases, but they are not discriminative enough for re-identification in uncontrolled environments. Thus, audio and visual signatures are separately learned and are then matched in a late fusion step depending on the person vs. sensor network localization (see section 3).

2.1 Audio Signature

Voice activity detection can be performed through multiple features, a review of them is described in [11]. To distinguish a speech segment from segments containing only noise, we exploit the temporal structure of speech, dominated by a characteristic energy modulation peak at about 4 Hz as detailed in [24]. Audio flux is then split into one-second-frames (whith overlapping) and filtered by a [2Hz-16Hz] bandpass. The segments with an energy modulation higher than a trained threshold are classified as speech, and the others as noise.

To construct text-independent speaker representation, we use Gaussian Mixture Models (GMMs), the most widely used approach, which has shown to perform well on numerous speech databases [19]. A GMM is

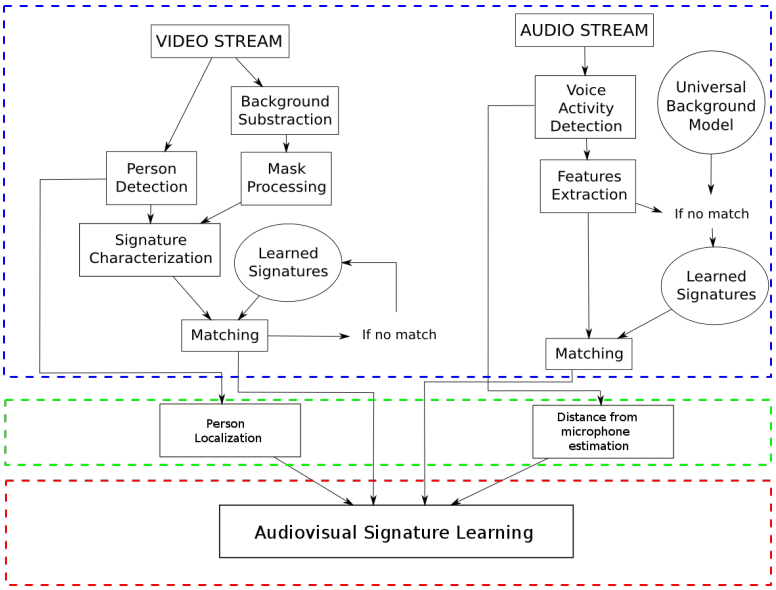

Figure 1: Synoptic of our multi-sensor platform

described as:

$$p(x|\lambda_{aud}) = \sum_{i=1}^{M} w_i g(x_i|\mu_i, \Sigma_i) \qquad (1)$$

where x is our feature vector and w_i is the weight of the Gaussian density component $g(x_i|\mu_i, \Sigma_i)$ with mean μ_i and covariance matrix Σ_i. Our audio signature for a speaker will be defined as:

$$\lambda_{aud} = \{w_i, \mu_i, \Sigma_i\} \qquad i = 1, ..., M. \qquad (2)$$

The features in vector x are based on a linear filter-bank of 19 filters derived cepstra, using perceptive MEL scale. Following the parametrization described in [3], each feature vector is composed of 50 coefficients, namely 19 static, 19 delta and 11 delta-delta and the delta energy.

Instead of computing the model from scratch, which requires a large amount of training data, we rather chose to train an Universal Background Model (UBM), using the BREF corpus [13] of 90 (50 females and 40 males) different speakers, for a total of 167,359 sentences, giving a prior, general representation of the average speaker. This model is estimated via the Expectation-Maximization (EM) algorithm [6] and adapted via a Maximum *A Posteriori* (MAP) estimation to obtain the desired speaker model with limited training data [20], usually 2-3 minutes per speaker are sufficient.

Considering a feature vector y of an observed speech segment Y, the similarity score is then computed as the log-likelihood ratio (LLR) between the hypothesis of Y spoken by our trained speaker vs. Y spoken by another one, closer to the UBM:

$$LLR(y) = log(p(y|\lambda_{GMM_x})) - log(p(y|\lambda_{UBM})) \qquad (3)$$

2.2 Visual Signature

Visual person re-identification has been a popular area of research in the past few years, involved in applications such as robotics, multimedia and in particular video surveillance from which many approaches emerged [1]. It raises several challenges, especially in Non-Overlapping Field Of View (NOFOV) camera network cases, such as partial occlusions, illumination changes, pose and viewpoint variations, changes in color response or unconstrained scenarii.

In our approach, the background subtraction, consisting in segmenting the regions belonging to the moving subject in a scene, is carried out by an online GMM clustering of the background [27].

For people detection, we use histograms of oriented gradient [5]. The person's shape is characterized by the distribution of local intensity gradients or edge directions. A linear SVM is then fed for person/non-person classification.

Once the mask of the foreground is applied on the detection, features are extracted to build discriminative descriptors of the person. Feature design has been widely explored and a recent review can be found in [22]. The majority of them follows a part-based model. In our case, the Symmetry-Driven Accumulation of Local Features (SDALF) introduced in [8], the silhouette is subdivided into head, torso and legs by looking for horizontal axis separating regions with strongly different appearance and similar area.

The SDALF descriptor achieves excellent state-of-the-art performances [8] by its robustness against very low resolution, occlusions and pose, viewpoint and illumination changes. It is then computed as a combination of the 3 following features extracted from the body

parts:

- HSV histograms, more robust to illumination changes than in RGB space by separating color and intensity and weighted by a gaussian kernel centered on their vertical axis of symmetry to emphasize central pixels, since the information they carry is more relevant.

- Maximally Stable Color Region (MSCR): this descriptor introduced in [9] computes color distances between pixels to find homogeneous areas and models them by elliptic blobs. The blobs are then represented by their area, centroid second moment matrix and color.

- Recurrent High-Structured Patches (RHSP), introduced in [8] also to characterize texture in high-entropy regions by analysing invariance of randomly extracted patches.

Figure 2 shows the Cumulative Match Curves (CMC) for the 3 features on 3 subsets of the ETHZ dataset [25] including a total of 8580 frames. All three features carry complementary information, yet HSV histograms perform satisfactorily by themselves in general cases, with a far lower computational cost than RHSP in particular. In the context of our application, a low CPU cost and a compact descriptor are required. Subsequently, MSCR and RHSP will then no more be considered.
To compute a storable and transportable visual signature, a k-means clustering is applied on a set of descriptors of a single person, then the k closest to each centroid are concatenated to form the signature, containing great variations of views and poses:

$$\lambda_{vid} = \{histo_i\} \qquad i = 1, ..., k. \tag{4}$$

This task discards redundant images, gaining compacity. The more different views from cameras the signature is confronted to, the higher value for k is required. In a single-camera case, with an inclination $\alpha = 25°$, $k = 6$ clusters provide satisfactory description of the appearance variabilities.
For the matching step, histograms similarities are computed by the Bhattacharyya distance [2].

3 Towards an Audiovisual Signature

As mentioned in the preamble, we merge the audio and video models in a late fusion step. Therefore, we look for environmental overlapping zones where the sensors are functional, thereby partitionning the space into zones of audio and/or video detection considering the relative positioning of the ambient cameras and microphones.

3.1 Audio Localization

Sound source localization in robotics is usually performed by exploiting binaural cues such as the Interaural Level Difference (ILD) and the Interaural Time Difference (ITD). As for human perception [18] these cues provide horizontal azimuth estimation, while elevation can be inferred through the filtering amount due to reflections in the pinna [23].
In our single-microphone context, however, such cues are impossible to use but acoustic features for close distance-to-listener estimation can be extracted. Beside naive descriptors such as sound intensity (there is a 6-dB loss in sound pressure per doubling the distance in free spaces), distance appears to be correlated to the Direct-to-Reverberant energy Ratio (DRR) [16]. A recent survey of cues for auditory distance perception in humans can be found in [12]. For synthetic audio signals, with known spectrum and temporal envelope, this ratio can be estimated through interaural cross-correlation, spectral variance, spectral envelope or buildup and decay analysis [14]. These cues fail for speech signals because of their non-stationary natures, therefore, we used a measure of intelligibility described in [7] as the Speech to Reverberation Modulation energy Ratio (SRMR).
The modulation spectrum of a dry speech signal (spectrum of its temporal envelope) has components distributed around 2-16Hz, with peaks at 4Hz, the common syllabic rate. Adding reverberation whitens the modulation spectrum and components in higher modulation frequency bands appear as shown in Figure 3.

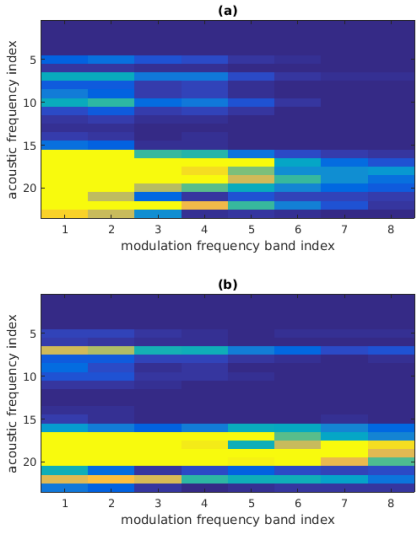

Figure 3: Modulation spectrum of the same frame for clean speech (a) and reverberated speech (b)

The yellow part on the modulation spectrum pictures its highest values. For clean signal in (a), one can

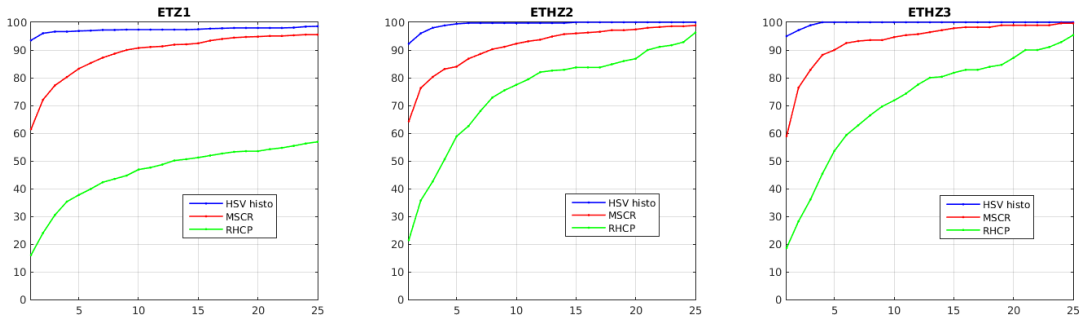

Figure 2: CMC curves of HSV histograms in blue, MSCR in red and RHSP in green for 3 subsets of ETHZ

observe that this modulation energy is restrained in the first 4 modulation frequency bands while it extends on the 8 modulation frequency bands for reverberated speech (see Table 1 for index/modulation frequencies band correspondence).
The input speech signal is first filtered by a 23-channel gammatone filter bank, then SRMR is defined as the ratio between the energy in low modulation frequency bands and the energy in high modulation frequency bands:

$$SRMR = \frac{\sum_{k=1}^{4} \overline{\varepsilon}_k}{\sum_{k=5}^{K*} \overline{\varepsilon}_k} \quad (5)$$

with $\overline{\varepsilon}_k$ the average modulation energy in band k.

Table 1: Modulation Filter Center Frequencies (f_c) and Bandwiths (BW) Expressed in Hz

	Modulation Frequency Band Index							
	1	2	3	4	5	6	7	8
f_c	4.0	6.5	10.7	17.6	28.9	47.5	78.1	128.0
BW	1.9	3.4	5.9	9.8	15.9	26.4	43.2	70.8

An updated version of the metric is described in [21] where text independence and pitch robustness are improved. It is observed that this metric is correlated to the distance as in Figure 4. The SRMR is displayed in blue and is computed every second from distances whose inverse values are display in red. The closer to the microphone (the peak at 50 seconds in our example), the higher the intelligibility is. SRMR will then be interpreted as a proximity confidence index.

3.2 Video Localization

Unlike audio percepts, the positions on the ground plane of the visual detections can be easily inferred from a camera view. We calibrate beforehand the ground plane relatively to the camera by using grids placed on the ground plane as shown in Figure 5.

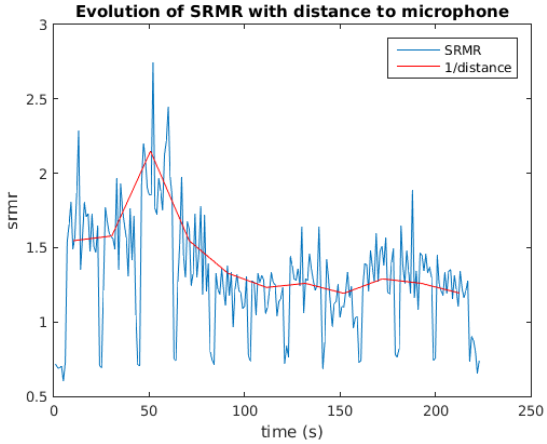

Figure 4: SRMR of 11 iterations of a speech segment from different distances and a frame length of 1s

3.3 Audiovisual Fusion

As introduced before, audiovisual fusion issue is tackled as a search of both modality salient areas at each instant. The room is equipped with several cameras C and microphones M in a way that a subset of microphones $\{M_K\}$ is in the field of view of at least one camera.
For a camera and a microphone M_k, let us consider $p_{aud}(t, M_k)$ and $p_{vid}(t, M_k)$ the models of the audio and video percepts respectively, at instant t. Their components are:

- the best match $\lambda_{aud,i}$ given by the speaker verification and the SRMR as the associated Audio Confidence Index (ACI);
- the best match $\lambda_{vid,j}$ given by the visual reidentification and the inverse of the euclidean distance to the microphone M_k as the associated Visual Confidence Index (VCI).

$$p_{aud}(t, M_k) = \begin{cases} \lambda_{aud,i} \\ ACI_{t,M_k} = SRMR_{t,M_k} \end{cases} \quad (6)$$

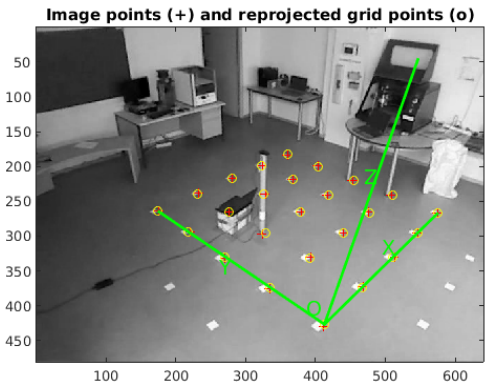

Figure 5: Extrinsic parameters estimation

$$p_{vid}(t, M_k) = \begin{cases} \lambda_{vid,j} \\ VCI_{t,M_k} = \frac{1}{\sqrt{(x-x_{Mk})^2+(y-y_{Mk})^2}} \end{cases} \quad (7)$$

Our joint measure for audiovisual saliency map is then defined as the following AudioVisual Confidence Index:

$$AVCI_{t,M_k} = ACI_{t,M_k} * VCI_{t,M_k} \quad (8)$$

For values of $AVCI_{t,M_k}$ greater than a threshold th, depending on how much confidence is permitted, audio and visual signatures are fused to define the detected person as the couple $(\lambda_{aud,i}, \lambda_{vid,j})$.
In multiple detection cases, each potential couple (i, j) at instant t is estimated by the probability:

$$p(\lambda_{av,ij}, t_1) = \frac{1}{N(t_1)} \quad (9)$$

with $N(t_1)$ the number of potential couples at instant t_1. For each potential couple (i, j) observed again at instant t_2, this probability is then updated:

$$p(\lambda_{av,ij}, t_1 + t_2) = p(\lambda_{av,ij}, t_1) + p(\lambda_{av,ij}, t_2) - p(\lambda_{av,ij}, t_1) * p(\lambda_{av,ij}, t_2) \quad (10)$$

Spatiotemporal analysis enables association ambiguity resolving and the most likely couples are then formed at the end of the training stage.

4 Experiments and Evaluations

4.1 Implementation

To the best of our knowledge, there is no public database matching the context of our problematic, therefore, we acquired audio and video data for our own evaluations. In a typical meeting room, of size 6m by 6m, 2 sensors are placed, a camera in one corner of the room, at height $h_{cam} = 2.5m$ and inclination $\alpha = 25°$ approximatively. An USB-microphone MXL-AC 404, generally used in video conferences, placed in the center of the room (in the field of vision).
The dataset is composed of 3 participants covering every location in the room while broadcasting 81 iterations of a clean speech segment from BREF, a large read-speech corpus in French [13] through a Bluetooth speaker. The data are then split into a training set of 486 speech segments with 544 visual frames with which the audiovisual locking zone is learned and a testing set of 243 speech segments with 222 visual farmes. The total duration of the dataset is 1 hour and 34 minutes.
To compute our audio signature, we used ALIZE, an open-source platform for speaker recognition [3] and its high level LIA RAL toolkit. The audio features are computed using the open source SPro toolkit. The UBM model size is composed of 512 Gaussian components (with diagonal covariance matrices). The HOG person detector, as well as the background substraction and the features extraction are computed using OpenCV library into a MATLAB environment.
Visual detections suffer from numerous false alarms, we add a post-processing filtering step. False detections induced by the background are recurrently observed. Nevertheless, they exhibit null foreground masks. Let us consider a subset of visual detections. Out of 302 detections, 182 are false alarms. Then, we filter them considering that the mask must include a minimum percentage p of pixels associated to detected mobile zones, discarding the others. Using another subset of annotated detections, we then trained a 1-D linear SVM on the sum of the foreground pixels to find the minimum percentage $p = 16\%$. 175 false detections are thus removed, the remaining 7 contain partial detections (torso or legs) with a great proportion of foreground pixels.
Figure 6 depicts the distribution of foreground pixels on samples of detection masks. Two clusters stand out pointedly, one for the false positives, circled in red and one another for the true positives, circled in green.

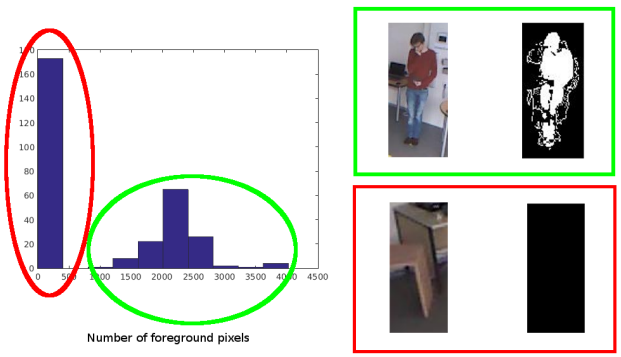

Figure 6: Histograms of mask pixels, in red false alarms, in green true positives

4.2 Evaluations and discussion

Audiovisual learning. For both visual and audio percepts, signatures of the 3 subjects are accurately trained, all the detections are correctly brought together among 3 distinct models. The representations are designed to be robust to far more challenging conditions than the ones of our protocol, hence their efficiency. Indeed, it contains a limited number of relatively dissimilar subjects, flagrant on Figure 7. This figure displays the thumbnails corresponding to the histograms composing the video signatures of the three participants, where inner variabilities of pose and illuminations are well described by the k-means clustering.

Figure 7: Thumbnails used for signature computation for the 3 participants.

AVCI values are computed from the audio and video percepts of the training data for the 3 participants. In blind spots of the camera, the Visual Confidence Index is set to zero and the AVCI is not computed at the location of the microphone (center), since the speaker and the microphone cannot stand on the same spot.
The audiovisual fusion zone is delimited by AVCI values greater than a predefined threshold th.
To tune this value, we fed a gaussian Kernel SVM with AVCI values of the training dataset and evaluated the classification error rate, function of th, shown in Figure 8.
The threshold is fixed at : $th = 0.4$, for which the classification error rate is lower than 10% for the 3 participants and the contour of the audiovisual fusion zone displayed in Figure 9 d).

Verification per both zone and ID. The audiovisual fusion zone is then confronted to our testing data. From the AVCI values computed on them and the learned threshold th, binary maps are extracted and shown in Figure 9 a), b) and c). The results of the joint ID/zone classification are shown in Table 2.

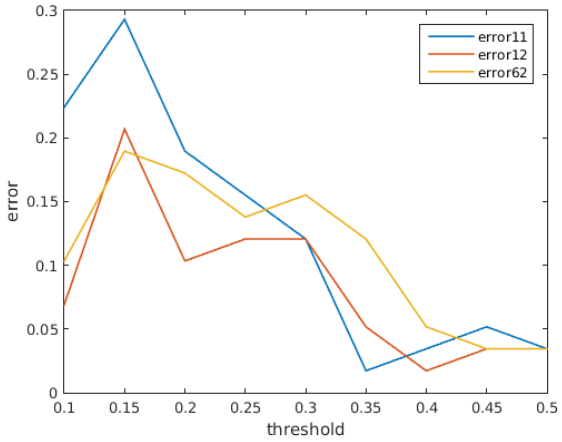

Figure 8: Classification Error Rate

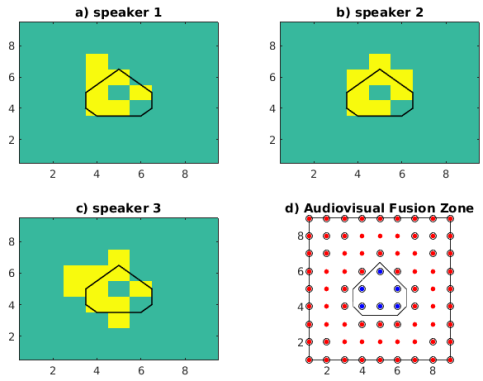

Figure 9: a), b) and c) : Binary maps of the AVCI for the 3 participants, d) fusion zone (blue)

The errors describe both false audiovisual associations and false zone estimations.

	Locations	Blind Spots	Classification Error Rate
$P1$	81	23	0.051
$P2$	81	23	0.069
$P3$	81	23	0.103

Table 2: Classification Results

5 Conclusion

In this paper, we presented a training method of an audiovisual signature of a person by coupling state-of-the-art approaches for the two modalities computed separately and merged lately.
The main contribution of this paper is the validation of a novel audiovisual confidence index to find saliency areas for both audio and visual percepts. This index is based on spatio-temporal coherence between visual

localization and auditive distance estimation at several instants of the training phase.
Future work will focus on spatiotemporal analysis of the audiovisual trajectories for fusion, then on transporting the presented audiovisual signature in a larger network of sparse sensors. With audiovisual signatures of multiple persons in a room, we plan to infer activity recognition in real-time for meetings, lectures or work groups from interaction between dominant participants.

References

[1] A. Bedagkar-Gala and S. K. Shah. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270 – 286, 2014.

[2] A. Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhya: The Indian Journal of Statistics (1933-1960)*, 7(4):401–406, 1946.

[3] J.-F. Bonastre, F. Wils, and S. Meignier. Alize, a free toolkit for speaker recognition. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 1, pages 737–740, March 2005.

[4] C. Busso, S. Hernanz, C.-W. Chu, S. il Kwon, S. Lee, P. Georgiou, I. Cohen, and S. Narayanan. Smart room: participant and speaker localization and identification. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 2, pages ii/1117–ii/1120 Vol. 2, March 2005.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893 vol. 1, June 2005.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[7] T. Falk, C. Zheng, and W.-Y. Chan. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(7):1766–1774, Sept 2010.

[8] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2360–2367, June 2010.

[9] P.-E. Forssen. Maximally stable colour regions for recognition and matching. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, June 2007.

[10] T. Germa, F. Lerasle, N. Ouadah, and V. Cadenat. Vision and {RFID} data fusion for tracking people in crowds by a mobile robot. *Computer Vision and Image Understanding*, 114(6):641 – 651, 2010. Special Issue on Multi-Camera and Multi-Modal Sensor Fusion.

[11] S. Graf, T. Herbig, M. Buck, and G. Schmidt. Features for voice activity detection: a comparative analysis. *EURASIP Journal on Advances in Signal Processing*, 2015(1):1–15, 2015.

[12] A. J. Kolarik, B. C. J. Moore, P. Zahorik, S. Cirstea, and S. Pardhan. Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. *Attention, Perception, & Psychophysics*, pages 1–23, 2015.

[13] L. F. Lamel, J. luc Gauvain, M. Eskenazi, and M. E. Limsi-cnrs. Bref, a large vocabulary spoken corpus for french. pages 505–508.

[14] E. Larsen, N. Iyer, C. R. Lansing, and A. S. Feng. On the minimum audible difference in direct-to-reverberant energy ratio. *The Journal of the Acoustical Society of America*, 124(1):450–461, 2008.

[15] R. Mazzon, S. F. Tahir, and A. Cavallaro. Person re-identification in crowd. *Pattern Recogn. Lett.*, 33(14):1828–1837, Oct. 2012.

[16] D. H. Mershon and L. E. King. Intensity and reverberation as factors in the auditory perception of egocentric distance. *Perception & Psychophysics*, 18(6):409–415.

[17] A. Mogelmose, C. Bahnsen, T. Moeslund, A. Clapes, and S. Escalera. Tri-modal person re-identification with rgb, depth and thermal features. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 301–307, June 2013.

[18] L. Rayleigh. On our perception of sound direction. *Philosophical Magazine Series 6*, 13(74):214–232, 1907.

[19] D. A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. *Speech Commun.*, 17(1-2):91–108, Aug. 1995.

[20] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(13):19 – 41, 2000.

[21] J. F. Santos, M. Senoussaoui, and T. H. Falk. An updated objective intelligibility estimation metric for normal hearing listeners under noise and reverberation. In *International Workshop on Acoustic Signal Enhancement (IWAENC)*, September 2014.

[22] R. Satta. Appearance descriptors for person re-identification: a comprehensive review. *CoRR*, 2013.

[23] A. Saxena and A. Ng. Learning sound location from a single microphone. In *Robotics and Automation, 2009. ICRA '09. IEEE International Conference on*, pages 1737–1742, May 2009.

[24] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 1331–1334 vol.2, 1997.

[25] W. Schwartz and L. Davis. Learning Discriminative Appearance-Based Models Using Partial Least Squares. In *Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing*, 2009.

[26] A. Tawari and M. Trivedi. Speech based emotion classification framework for driver assistance system. In *Intelligent Vehicles Symposium (IV), 2010 IEEE*, pages 174–178, June 2010.

[27] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, pages 28–31 Vol.2, Aug 2004.