On-line cross-modal adaptation for audio-visual person identification with wearable cameras

Alessio Brutti, Andrea Cavallaro

***Abstract*—We propose an audio-visual target identification approach for egocentric data with cross-modal model adaptation. The proposed approach blindly and iteratively adapts the time-dependent models of each modality to varying target appearance and environmental conditions using the posterior of the other modality. The adaptation is unsupervised and performed on-line, thus models can be improved as new unlabelled data become available. In particular, accurate models do not deteriorate when a modality is underperforming because of an appropriate selection of the parameters in the adaptation. Importantly, unlike traditional audio-visual integration methods, the proposed approach is also useful for temporal intervals during which only one modality is available or when different modalities are used for different tasks. We evaluate the proposed method in an end-to-end multi-modal person identification application with two challenging real-world datasets and show that the proposed approach successfully adapts models in presence of mild mismatch. We also show that the proposed approach is beneficial to other multi-modal score fusion algorithms.**

***Index Terms*—model adaptation, multimedia systems, person identification, wearable cameras**

I. Introduction

The increasing availability of body-worn cameras is facilitating applications such as life-logging and activity detection [42], [68]. In particular, recognising objects or the identity of humans from egocentric data is an important capability that enables tasks such as diarization and interaction recognition. Typically, a model of the target (e.g. a person) is first acquired during a short enrolment stage. Next, the identity of the target should be verified using audio and video signals during future interactions (i.e. when the target appears again in front of the worn camera). While mono-modal audio or video systems are successful in controlled scenarios [12], [38], [67], constraints on the environment, on the target position or appearance, and on the speech signals need to be relaxed to enable applications with body-worn cameras. To this end, we aim to exploit the complementarity of the observations generated by multiple sensing modalities [63].

Multi-modal person identification is of interest for surveillance [67], meeting analysis [65] and biometrics [7]. These applications generally assume the availability of high quality (or at least frontal) views of the subjects, jointly with a speech signal. These assumptions are too restrictive for body-worn cameras as new challenges arise when addressing the egocentric person (re-)identification problem. For example, the amount of training and enrolment data is generally limited and no large datasets exist to train on for a variety of targets that could appear under varying scene conditions. Ideally, as soon as a new target appears before the wearable camera the system should be able to generate reliable person models.

Because the target appearance *and* the environment change over time, it is generally not possible to collect enough samples to train models that account for these changes. Moreover, retraining models each time the conditions change may be computationally infeasible. For example, batch Maximum A-Posteriori (MAP) adaptation of models [58] cannot be used because of the unavailability of data for all targets. A possible solution is to continuously adapt the models to changes in target pose, reverberation, background noise and illumination. Unsupervised adaptation of target models has been addressed for a single modality in speaker verification [11], [37], [54] and in domain adaptation for visual person re-identification [45], [70]. However, to avoid performance deterioration, conservative adaptation strategies are needed when using only one modality.

To address these problems, we propose an on-line continuous and multi-modal target-model adaptation framework. We define time-dependent target models that are adapted in an unsupervised fashion by exploiting the complementarity of multiple modalities as soon as a new observation is available. The proposed on-line adaptation addresses the problem of model mismatch due to varying environmental conditions and changing target appearance. Moreover, the enrolment set is expanded thus increasing the overall model accuracy even in low mismatch conditions. Finally, the proposed approach allows us to control the speed of model adaptation and to cope with situations when one of the modalities is underperforming. In summary, our main contributions are the following. 1) We address the multi-modal target identification task for egocentric applications with wearable audio-visual devices. These applications are characterised by novel challenges caused by varying environmental conditions, a variety of unusual viewpoints, image and sound distortions, potential unavailability (or unreliability) of one of the modalities and limited amount of training data. 2) We introduce temporal models that are continuously adapted in an unsupervised manner using information from another modality. The proposed adaptation approach can complement traditional multi-modal integration algorithms based on score fusion or joint classification. 3) The proposed model adaptation is effective not only for multi-modal (audio-visual) identification tasks, but also for improving the performance of each mono-modal model. Therefore the proposed approach is also beneficial when only one of the two modalities is available (e.g. a person is silent before the

A. Brutti is with the Center for Information and Communication Technology, Fondazione Bruno Kessler, 38123, Trento, Italy (e-mail: brutti@fbk.eu). This work was done when the first author was visiting Queen Mary, University of London.

A. Cavallaro is with the Centre for Intelligent Sensing, Queen Mary University London, E1 4NS, London, UK (e-mail: a.cavallaro@qmul.ac.uk)

camera or talks from outside the field of view) and fusion is not applicable. Furthermore, as long as the sensors capture the same scene, the proposed method is applicable also when each modality is used for a different task, such as speech recognition and target tracking.

The paper is organised as follows. Section II presents the related state-of-the-art and highlights the main approaches for joint audio-video processing. The problem we address is defined in Section III. Section IV describes the proposed solution. Section V presents an implementation example for joint audio-visual target identification. Section VI discusses the experimental results. Finally, Section VII concludes the paper with our final remarks and the description of future work.

II. Related Work

Audio-visual recognition approaches exploit the complementarity of two data streams and can be divided in three main categories: early, mid and late fusion methods [43]. Moreover, hybrid fusion methods combine at least two of the above-mentioned categories (see Table I).

Early fusion methods combine audio and video features before processing the joint feature set. For example, feature vectors may be stacked and then processed jointly for interaction recognition using features derived from spatial cues. Stacking audio and video features in combination with a Hidden Markov Model (HMM) yields better results than processing the two modalities independently [48]. The dimensionality of the stacked feature vector can be reduced through Principal Component Analysis and Independent Component Analysis to remove redundant data across the modalities [64].

Late fusion methods process audio and video independently, and then combine the final mono-modal scores or decisions. This combination is often weighted by the reliability of each modality [28], [32], [44]. Reliability measures can relate to the signals (e.g. SNR), to the models [32] or to the recognition rate of each classifier by learning appropriate weights [44]. Decision selection is also a late fusion strategy that is typically based on the reliability or the discriminative power of each expert [18], [29], [47], [60]. Examples include articulate hard-coded decision cascades driven by reliability [29] and adaptive weights based on an estimation of the model mismatch from the score distributions [60]. The final score combination may also consist of a further classification stage based on a Gaussian Mixture Model (GMM), Support Vector Machine or Multi-Layer Perceptron [6], [59]. In this case, scores are stacked and treated as feature vectors in order to increase target discriminability. Confidence measures can also be included to improve classification [4], [13]. The main drawback is the need for a further training stage. Late fusion methods are generally efficient and modular, as other modalities or sub-systems can be easily added.

Mid-fusion methods process features independently and merge the modalities in a joint classification stage. These methods are used for activity diarization or recognition and typically employ a Multi-Stream HMM [43], [69] or Dynamic Bayesian Networks [53]. In general, a weighted sum of the log-likelihoods is adopted, which is equivalent to late fusion when, instead of HMMs, a time-independent classifier is used. In [62] the stream with the highest posterior is used in combination with an HMM for Audio-Visual Speech Recognition. Asynchronous HMM combinations are needed when the sampling rate of each modality is different and misalignments between the time instants when a target manifests itself in each modality occur.

Finally, *hybrid methods* perform fusion in at least two stages of their pipeline. As an example, in Co-EM [14] models are iteratively updated in the maximisation step, exchanging labels between multiple views of the same dataset and minimising the disagreement between modalities. Examples of Co-EM in speech recognition or multi-modal interaction are referred to as co-training [40] and co-adaptation [19]. *Co-training* can be used for traffic analysis using multiple cameras [40]. The goal is to generate a larger training set for batch adaptation through unsupervised labelling of unseen training data. This labelling is performed based on the agreement between weak mono-modal classifiers trained on small labelled datasets. *Co-adaptation* for audio-visual speech recognition and gesture recognition jointly adapts audio and visual models using unseen unlabelled data of the new application domain by maximising their agreement [19]. While the underlying idea of maximising the agreement of multiple classifiers to label unseen data is similar to our approach, state-of-the-art methods are not suitable for adapting the time-varying models that are of interest in this work. In fact, multi-modality is used in the development phase of the system to allow the use of larger training datasets whose manual annotation is impractical.

III. Problem Formulation

Let us consider S uncooperative targets (to be) enrolled in the system. The value of S is unknown and time-varying. Audio-visual target models are acquired on-the-fly during enrolment. Let K_t be an audio-visual segment consisting of a set of audio samples x_t and images $\mathcal{I}_t$ where the target manifests itself in both modalities simultaneously. For example, when the target is a person, the enrolment happens when the target talks in front of the camera.

Let w_t^i be the feature vector of K_t for modality i and $\mathbf{w}_t^i = [w_t^i, \ldots, w_{t+T-1}^i]$ be the observation vector, where T is the duration of the observation window (i.e. the number of observations). Finally, let $\mathbf{\Theta}^i = \left[\Theta^i(1), \ldots, \Theta^i(S)\right]$ be the model parameters for modality i. The model set $\hat{\mathbf{\Theta}}^i$ can be estimated through MAP adaptation as:

$$\hat{\mathbf{\Theta}}^i = \arg\max_{\mathbf{\Theta}^i} p(\mathbf{w}_t^i|\mathbf{\Theta}^i) g(\mathbf{\Theta}^i), \tag{1}$$

where $p(\mathbf{w}_t^i|\mathbf{\Theta}^i)$ is the probability of observation given the model and $g(\mathbf{\Theta}^i)$ is the prior for the model distribution. Since closed-form solutions of eq. 1 are in general not available, Expectation Maximization (EM) is typically employed and the parameters of the prior $g(\mathbf{\Theta}^i)$ are estimated on a large dataset containing a variety of targets [58]. However, traditional MAP adaptation schemes are not suitable for our application scenario because target models need to evolve over time to adapt to changes in the (multi-modal) target appearance as well as in the environment. Moreover, collecting

TABLE I: Comparison of state-of-the-art methods for late, mid, early and hybrid fusion. KEY – Ref.: reference; Prop.: proposed method; MFCC: Mel-Frequency Cepstral Coefficients; i-Vectors: see Sec. V-A; DCT: Discrete Cosine Transform; RGB: Red, Green and Blue color channels; GMM: Gaussian Mixture Models; SVM: Support Vector Machine; MLP: Multi-Layer Perceptron; PCA: Principal Component Analysis; LFA: Local Feature Analysis; HMM: Hidden Markov Models; MS-HMM: Multi-Stream Hidden Markov Models; SIFT: Scale Invariant Feature Transform; DBN: Dynamic Bayesian Network; LBP: Local Binary Pattern; PLDA: Probabilistic Linear Discriminant Analysis; ROI: Region Of Interest; ICA: Independent Component Analysis; LDA: Linear Discriminant Analysis; conf.: confidence; norm.: normalisation; Eucl.: Euclidean; Batt.: Bhattacharyya; Dist.: distance; n.a.: not available; *: not for identification.

Ref.	**Fusion**	**Features**		**Classifier**		**Method**	**Data**
		Audio	Video	Audio	Video		
[13]	Late	MFCC	RGB + gray level	GMM	MLP	score classification with conf. measures	XM2VTS [49]
[60]	Late	MFCC	PCA on gray level	GMM	GMM	adaptive weighted sum with conf. measures	VidTIMIT [61]
[18]	Late	MFCC	DCT	GMM	GMM	reliability based fusion	own data
[6]	Late	MFCC	PCA	GMM	Eucl. Dist.	GMM	ViBE [66]
[29]	Late	MFCC	PCA	HMM	Eucl. Dist.	decision cascade, conf. measures	MVGL-AVD [30]
[44]	Late	MFCC	PCA	GMM	Eucl. Dist.	matcher weighting	XM2VTS [49]
[28]	Late	MFCC	DCT	GMM	Eucl. Dist.	adaptive weighted sum	CLEAR 2007 [50]
[36]	Late	MFCC	ROI + gray level, lips + PCA	GMM	SVM	linear weight trained on dev. data	own mobile devices and AV-TIMIT [35]
[32]	Late	MFCC	DCT, LFA	HMM	FaceIt [9]	score norm., reliability, weighted sum	XM2VTS [49]
[47]	Late	i-Vectors	LBP Hist.	PLDA	Hist. Dist.	score linear regression	Mobio [47]
[4]	Late	MFCC	ROI	GMM	GMM	confidence based sum rule	VidTimit [61] and AusTalk [16]
[62]*	Mid	MFCC	Mouth contour	HMM		Maximum stream posterior	ASR on own data
[69]*	Mid	Pitch, energy, rate, spatial coherence	RGB blob of head and right hand	MS-HMM		asynch. MS-HMM for action recognition	meetings
[53]*	Mid	MFCC	SIFT	DBN		features from audio, video and joint spaces for diarization	TRECVID [1] and AMI [17]
[43]	Mid	MFCC	ROI (mouth)	MS-HMM		PCA + LDA, asynch. MS-HMM	M2VTS [49]
[64]*	Early	DCT	intensity	n.a.		PCA + ICA, no classification, theoretical study	event recognition
[48]*	Early	local coherence, keyword confidence	RGB based head blob	HMM		stack of spatial feature vectors, interaction recognition based on spatial information	meeting
[40]*	Hybrid	n.a.	gray level	n.a.	Eucl. Dist.	co-training of multiple camera views, no audio	traffic
[19]	Hybrid	MFCC	head movement	GMM	GMM	Batch co-adaptation, no fusion	Audio-video speech and gesture recognition
Prop.	Hybrid	i-Vectors	RGB	Cosine Scoring	Batt. Dist.	Multi-modal Cross-adaptation	body-worn, seminars

a sufficient amount of adaptation material may be unfeasible for some targets. Therefore, the models need to quickly adapt to new conditions, potentially using only a single in-domain observation (i.e. $T = 1$).

IV. On-line multi-modal adaptation

To temporally adapt the models using unlabelled new observations we propose to use a Kalman Filter whose gain depends on the posterior probability of the complementary modality. We follow the terminology of the speaker verification community and derive the Kalman Filter equations from the EM solution of the MAP adaptation problem [34], as discussed next.

A. Model adaptation

Let the dynamics of the time-varying target appearance be modelled as a first order Markov Chain. Therefore the model $\mathbf{\Theta}_t^i = \left[\Theta_t^i(1), \ldots, \Theta_t^i(S)\right]$ at segment t depends only on $\mathbf{\Theta}_{t-1}^i$ [52]:

$$g(\mathbf{\Theta}_t^i) = p(\mathbf{\Theta}_t^i | \mathbf{\Theta}_{t-1}^i). \tag{2}$$

In this case, we can reformulate in an on-line fashion the batch MAP adaptation problem of eq. 1, where for each new observation vector, w_t^i, we derive:

$$\hat{\mathbf{\Theta}}_t^i = \arg\max_{\mathbf{\Theta}^i} p(w_t^i | \mathbf{\Theta}_t^i) p(\mathbf{\Theta}_t^i | \mathbf{\Theta}_{t-1}^i). \tag{3}$$

Unless the distributions have special shapes, a close-form solution of this problem is in general not available. For this reason we resort to the EM algorithm [24]. EM performs two steps iteratively, namely the expectation step (E-step) and the maximization step (M-step). In the E-step, the expectation of the log-likelihood is evaluated on the observed data using the current parameter estimate. In the M-step, a new set of model parameters is derived by maximising the expected log-likelihood estimated in the E-step.

The EM solution of the Maximum Likelihood (ML) model estimation problem relies on the maximisation of the auxiliary function $Q(\mathbf{\Theta}_t^i, \mathbf{\Theta}_{t-1}^i)$ [33]:

$$Q(\mathbf{\Theta}_t^i, \mathbf{\Theta}_{t-1}^i) = E\left[\log p(\mathbf{\Theta}_t^i | w_t^i) | s, \mathbf{\Theta}_{t-1}^i\right] \tag{4}$$

$$= \textstyle\sum_{s=1}^{S} p(s | w_t^i, \mathbf{\Theta}_{t-1}^i) \log p(w_t^i | \mathbf{\Theta}_t^i, s), \tag{5}$$

where the latent variable $s = 1, \ldots, S$ represents the unknown identity of the target associated to the observation vector w_t^i. Similarly, when models are estimated in the MAP sense, the following auxiliary function is used [33], [34]:

$$R(\mathbf{\Theta}_t^i, \mathbf{\Theta}_{t-1}^i) = Q(\mathbf{\Theta}_t^i, \mathbf{\Theta}_{t-1}^i) + \log p(\mathbf{\Theta}_t^i | \mathbf{\Theta}_{t-1}^i), \tag{6}$$

where the term $\log p(\mathbf{\Theta}_t^i | \mathbf{\Theta}_{t-1}^i)$ accounts for the prior. We assume that vector models are independent of each other, so that probabilities factorise as follows:

$$p(w_t^i | \mathbf{\Theta}_t^i) = \prod_{s=1}^{S} p(w_t^i | \Theta_t^i(s)). \tag{7}$$

Without prior information about the statistics of the observation vector, we assume w_t^i to be normally distributed:

$$p(w_t^i|\Theta_t^i(s)) = B_w \exp\left(-\frac{\|w_t^i - \Theta_t^i(s)\|^2}{2{\sigma_w^i}^2}\right), \quad (8)$$

where ${\sigma_w^i}^2$ is the variance of the observation vector and B_w is the normalisation term independent of $\Theta_t^i(s)$. Similarly, we model the prior as normal distribution:

$$p(\mathbf{\Theta}_t^i|\mathbf{\Theta}_{t-1}^i) = \prod_{s=1}^{S} B_\Theta \exp\left(-\frac{\|\Theta_t^i(s) - \Theta_{t-1}^i(s)\|^2}{2{\sigma_\Theta^i}^2}\right), \quad (9)$$

where B_Θ is a constant normalisation term and ${\sigma_\Theta^i}^2$ is the variance of the model that accounts for the reliability of the prior (i.e. how different a new model can be from the previous one). Finally, we assume that the components of the covariance matrices are independent (i.e. the matrices are diagonal) and are independent of s and t. The latter assumption makes the problem mathematically more tractable.

The auxiliary function in eq. 6 therefore becomes:

$$\begin{aligned} R(\mathbf{\Theta}_t^i, \mathbf{\Theta}_{t-1}^i) &= \sum_{s=1}^{S} \pi_t^i(s) \log p\left(w_t^i|\Theta_t^i(s)\right) \\ &+ \sum_{s=1}^{S} \log p\left(\Theta_t^i(s)|\Theta_{t-1}^i(s)\right). \end{aligned} \quad (10)$$

where $\pi_t^i(s) = p\left(s|w_t^i, \Theta_{t-1}^i(s)\right)$ is the membership probability of target s. Note that in the traditional EM algorithm $\pi_t^i(s)$ is estimated in the E-step, performing a probabilistic association to each model component.

Weak models are difficult to improve and, when models are inaccurate, the membership probability will also be inaccurate. Therefore observations may be associated to the wrong target thus leading to a further deterioration of the models. To avoid this performance deterioration, one could use a conservative adaptation or acceptance thresholds on the posteriors so that the observation is used for adaptation only when the posterior exceeds a certain threshold [11], [37], [54]. To address this problem we instead introduce in eq. 10 the posterior probability of a complementary modality.

B. Cross-modal membership probabilities

Let j denote a complementary modality to i. We introduce j in eq. 10 by replacing the membership probability $\pi_t^i(s)$ with the posterior probability of the complementary modality $\pi_t^j(s)$:

$$\pi_t^j(s) = p(s|w_t^j, \Theta_{t-1}^j) = \frac{p(w_t^j|\Theta_{t-1}^j(s))}{\sum_{s=1}^{S} p(w_t^j|\Theta_{t-1}^j(s))}. \quad (11)$$

Expanding the logarithms and removing constant terms independent of $\mathbf{\Theta}_t^i$, eq. 6 can be written as:

$$\begin{aligned} R(\mathbf{\Theta}_t^i, \mathbf{\Theta}_{t-1}^i) &= -\frac{1}{2{\sigma_w^i}^2} \sum_{s=1}^{S} \pi_t^j(s) \|w_t^i(s) - \Theta_t^i(s)\|^2 \\ &- \frac{1}{2{\sigma_\Theta^i}^2} \sum_{s=1}^{S} \|\Theta_t^i(s) - \Theta_{t-1}^i(s)\|^2. \end{aligned} \quad (12)$$

If we equal to zero the derivative of $R(\mathbf{\Theta}_t^i, \mathbf{\Theta}_{t-1}^i)$ with respect to each model vector $\Theta_t^i(s)$:

$$\frac{\partial R(\mathbf{\Theta}_t^i, \mathbf{\Theta}_{t-1}^i)}{\partial \Theta_t^i(s)} = 0, \quad (13)$$

we obtain:

$$-\frac{\pi_t^j(s)}{{\sigma_w^i}^2} w_t^i + \left(\frac{\pi_t^j(s)}{{\sigma_w^i}^2} - \frac{1}{{\sigma_\Theta^i}^2}\right) \Theta_t^i(s) + \frac{1}{{\sigma_\Theta^i}^2} \Theta_{t-1}^i(s) = 0, \quad (14)$$

which leads to the Kalman Filter equations:

$$\hat{\Theta}_t^i(s) = (1 - \alpha_t^j)\hat{\Theta}_{t-1}^i(s) + \alpha_t^j w^i(t), \quad (15)$$

where:

$$\alpha_t^j = \frac{\pi_t^j(s)}{\pi_t^j(s) + \frac{{\sigma_w^i}^2}{{\sigma_\Theta^i}^2}}. \quad (16)$$

Note that eqs. 15 and 16 are the result of the Gaussianity assumption for the observations, the use of a single observation for model adaptation and the use of the likelihood of another modality instead of the expectation step. If different assumptions are adopted when solving the maximisation of the auxiliary function, this approach will not necessary lead to the same Kalman Filter equations.

For any target s we have $\alpha_t^j > 0$, even if very small. Therefore, the models of all targets $s' \neq s$ would be partially affected by the observation vectors generated over time by target s, if the target is persistent in the scene. This problem is due to the probabilistic association of eq. 11, which is not an issue in batch adaptation because observations are available for most of the models and therefore low-probability faulty associations are negligible. To address this issue in online adaptation, we employ a ML association between the observation vector and the model to adapt [34]. Therefore only the most likely model in the other feature domain j is updated. The new expected posterior becomes:

$$\pi_t^j(s) = \begin{cases} \frac{p(w_t^j|\Theta_{t-1}^j(s))}{\sum_{s=1}^{S} p(w_t^j|\Theta_{t-1}^j(s))} & \text{if } s = \hat{s}_{\text{ML}}^j \\ 0 & \text{otherwise} \end{cases}, \quad (17)$$

where:

$$\hat{s}_{\text{ML}}^j = \arg\max_{s=\{1,\ldots,S\}} p(w_t^j|\Theta_{t-1}^j(s)).$$

C. Discussion

The above adaptation is applied iteratively for each new observation vector. The term $\frac{{\sigma_w^i}^2}{{\sigma_\Theta^i}^2}$ is equivalent to the relevance factor [58] and controls how much the model is adapted using a new observation. In adaptation of mixture models, the relevance factor can be interpreted as the minimum number of observations associated to a given component and to be used to update its parameters. Instead, in eq. 16 the relevance factor is related to the uncertainty of the observation vector and the prior. When the prior is tight (σ_Θ^i is small) and the observations have high variance (σ_w^i is large), then $\alpha \simeq 1$ and the models are not adapted. If a measure of $\frac{{\sigma_w^i}^2}{{\sigma_\Theta^i}^2}$ is available, for example based on the degree of mismatch, it can be used to dynamically control model adaptation.

When the models of all modalities are accurate, the adaptation is equivalent to a traditional mono-modal MAP adaptation and it makes the models more robust injecting new in-domain training data [11]. When the models of one modality are inaccurate and those of other modalities are good, the inaccurate models are improved without affecting the good ones.

Let us consider two modalities i and j for target s and assume that the models of modality i are accurate: $p(w_t^i|\Theta_t^i(s)) > p(w_t^i|\Theta_t^i(s'))\ \forall s' \neq s$, whereas the models of modality j are inaccurate: $\exists s' : p(w_t^j|\Theta_t^j(s)) < p(w_t^j|\Theta_t^j(s'))$. In this case, model $\Theta_t^i(s')$ is erroneously updated using w_t^i. Let us assume that the posterior is linearly related to the models and test vectors. For instance, this assumption holds when cosine scoring is used in the classification stage or when the scoring function can be locally linearised. The posterior of the updated model of s' at time $t+1$ becomes:

$$\begin{aligned} & p(w_{t+1}^i|\Theta_t^i(s')) = p(w_t^i|(1-\alpha_t^j)\Theta_t^i(s') + \alpha_t^j w_t^i) \\ = \quad & (1-\alpha_t^j)p(w_t^i|\Theta_{t-1}^i(s')) + \alpha_t^j p(w_t^i|\Theta_t^i(s)). \end{aligned} \tag{18}$$

The difference $\Delta_t^i(s,s') = p(w_t^i|\Theta_t^i(s)) - p(w_t^i|\Theta_t^i(s'))$ between the models is then:

$$\Delta_{t+1}^i(s,s') = (1-\alpha_t^j)\Delta_t^i(s,s'). \tag{19}$$

Since $\alpha_t^j < 1$, the new difference between the models is smaller. When the models have similar probabilities (for example under strong environmental noise) this erroneous adaptation could be detrimental because models would get even closer. However, if $p(w_t^i|\Theta_{t-1}^i(s)) \gg p(w_t^i|\Theta_{t-1}^i(s'))\ \forall s' \neq s$, the ranking between the models is not compromised and the faulty modality j is corrected. Moreover, the membership probability $\pi_t^j(s)$ is expected to be low (i.e. all models would have similar probability) thus resulting in a very small adaptation. Finally, since models are inaccurate, the variance of the observation σ_w^i is likely to be high. If this measure is available or can be estimated from the data, the adaptation rate can be reduced by increasing the relevance factor in eq. 16 thus preventing a faulty adaptation.

Note that eqs. 15 and 16 are similar to the traditional mono-modal MAP-EM maximisation equations [33], but with the membership probability estimated from the complementary modality (i.e. without expectation step). Eq. 15 can be also seen as a particular case of the generalised Co-EM where only the disagreement between the modalities is minimised [14].

Figure 1 shows the block diagram of the proposed multi-modal model adaptation approach and the final fusion of the scores when $J \geq 2$ modalities are available. In the next section we present a specific example of this incremental unsupervised model adaptation.

V. Application to audio-visual identification

We now present an audio-visual multi-modal model adaptation implementation with cross-modal feedback of scores (Figure 2). We discuss the audio and visual feature extraction front-ends, the cross-modal adaptation and the score combination.

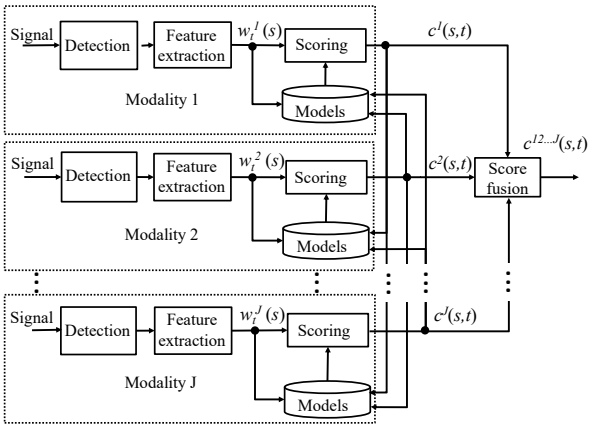


Fig. 1: Block diagram of the proposed multi-modal framework with J modalities, with feedback for model adaptation.

A. Feature extraction

Let the superscripts a and v denote the audio and the video modality, respectively. We aim to generate feature vectors to be used for enrolment and identification. Both model and test vectors are length normalised.

In the audio front-end, Mel-Frequency Cepstral Coefficients (MFCC) features are extracted on windows of 20 ms with 10 ms steps. A 30-dimensional feature vector is built with 15 Mel coefficients and their first derivatives. We use the Total Variability (TV) framework [23], [21] for audio feature extraction because of its robustness and flexibility [2]. The TV approach uses a data-driven feature extractor based on factor analysis to map the sequence of MFCC feature vectors into a low-dimensional vector, the i-Vector. The i-Vector, w_t^a, encodes speaker-related factors, i.e. the speaker representation in the speaker subspace.

The MFCC sequence, $\mathbf{x}_t$, of a given segment, K_t, can be represented as the supervector[1] M of the corresponding GMM, obtained through MAP adaptation from the Universal Background Model, $\mathbf{U}$. The dimensionality of the supervector M is high as it is the product of the number of features and the number of Gaussian components. With the TV approach a high-dimensional sentence-dependent vector M can be represented as:

$$M = \mu + Tw_t^a, \tag{20}$$

where μ, the supervector of $\mathbf{U}$, captures speaker-independent factors; and T is the TV matrix, estimated on a specific, possibly in-domain, training set.

When multiple enrolment segments are available, we obtain the model $\Theta^a(s)$ of target s as the average of the i-Vectors:

$$\Theta^a(s) = \frac{1}{N_s}\sum_{t=1}^{N_s} w_t^a(s), \tag{21}$$

where N_s is the number of enrolment segments. Given a test i-Vector w_t^a, we obtain the verification score (i.e. how likely

[1] A Gaussian supervector is the vector resulting from the concatenation of the mean vectors of a GMM.

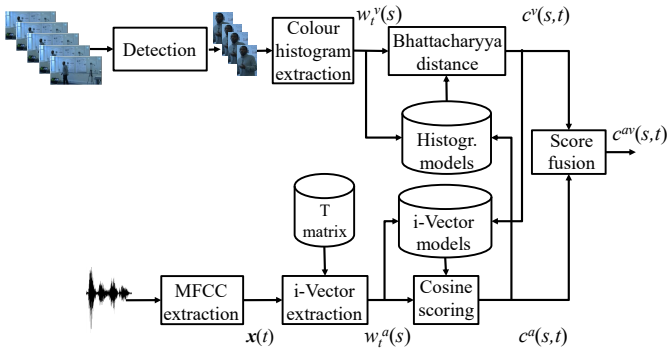


Fig. 2: Instantiation of the proposed multi-modal framework for audio-visual target recognition.

it is that the speech in segment t has been pronounced by speaker s) for each model through cosine scoring:

$$c^a(s,t) =< \Theta^a(s), w_t^a >, \tag{22}$$

which is a suitable metric for the TV framework [22].

In the video front-end, features are based on the colour histograms of the upper body of the target, which is estimated using an Aggregate Channel Features (ACF) detector [25], [51]. The feature vector is obtained by concatenating the RGB histograms of nine non-overlapping regions. We assume that these histograms are stationary across a few consecutive frames.

The set of images $\mathcal{I}_t$ of segment t produces a sequence of feature vectors $\mathbf{f}_t = [f_t^T(1), f_t^T(2), \ldots, f_t^T(N_I)]^T$, where N_I is the number of images in each segment and ^{T} indicates the transpose operation. Following the same strategy adopted for the audio modality, each segment is represented with a single feature vector obtained as a column-wise averaging of $\mathbf{f}_t$:

$$w_t^v = \frac{1}{N_I} \sum_{m=1}^{N_I} f_t(m). \tag{23}$$

When multiple enrolment segments are available for a given target, we obtain the final model vector as the average of all the segments:

$$\Theta^v(s) = \frac{1}{N_s} \sum_{t=1}^{N_s} w_t^v(s). \tag{24}$$

During testing, the vector w_t^v is derived with the same procedure for each trial segment t. The similarity between w_t^v and a model vector $\Theta^v(s)$ is based on the Bhattacharyya Distance (BD) [31], [46], [56]:

$$c^v(s,t) = \text{BD}(\Theta^v(s), w_t^v) \tag{25}$$

$$= < \sqrt{\Theta^v(s)}, \sqrt{w_t^v} > . \tag{26}$$

The audio and video modalities are then integrated in two stages, namely model adaptation and score combination, as discussed below.

B. Multi-modal model adaptation and late score fusion

As for the multi-modal model adaptation described in Section IV, the probability $p(w_t^j|\Theta_{t-1}^j(s))$ of the other modality j for segment K_t is derived directly from the classification scores $c^j(s,t)$. Hence, eq. 17 becomes:

$$\pi_t^j(s) = \begin{cases} \frac{c^j(s,t)}{\sum_{s=1}^{S} c^j(s,t)} & \text{if } s = \arg\max_{s'=\{1,\ldots,S\}} c^j(s',t) \\ 0 & \text{otherwise} \end{cases}. \tag{27}$$

Note that while the scores resulting from the cosine distance used in the audio modality are between -1 and 1, given the high dimensionality of the feature vector the negative values are very small and can be rounded to zero.

Finally, we perform a late combination of the audio and video scores using the sum rule [39]:

$$c^{av}(s,t) = \gamma_t c^a(s,t) + (1-\gamma_t) c^v(s,t). \tag{28}$$

Several approaches exist for deriving the weight γ_t (see Sec. II and Table I), such as using a further classification stage [6], [13], [59]. However, while this approach is robust, it requires specific training data that are not available in our application scenario. Moreover, the weights should be continuously adapted to the varying conditions.

We relate the weights to confidence measures of each modality. The weights can be derived from the individual classification scores [29], [32], which quantify the matching of the model to the operational conditions. We found experimentally that a good estimation can be derived from the reciprocal of the variance of the classification scores, excluding the highest one. The idea is that if a modality is reliable the scores of all the targets but the highest one will be low and similar. Taking the reciprocal will therefore give a high weight to that modality. Thus, we define the weight γ_t as:

$$\gamma_t = \frac{\xi_t^a}{\xi_t^a + \xi_t^v}, \tag{29}$$

with

$$\xi_t^j = \frac{1}{\frac{1}{S-2} \sum_{s \neq s_{\text{ML}}^j} \left(c^j(s,t) - \mu_t^j \right)^2}, \tag{30}$$

where μ_t^j is the mean of the scores excluding the highest one and $j = \{a, v\}$.

To conclude, only when the combined score $c^{av}(s,t)$ exceeds the threshold τ the target identity s is accepted. Typically, τ is experimentally determined to achieve a desirable trade-off between false alarm rate and miss detection rate.

VI. Experimental Analysis

We validate the proposed approach by evaluating the mono-modal and multi-modal ("Late") systems *with* and *without* model adaptation. Moreover, we discuss the benefits of applying the proposed model adaptation strategy to two state-of-the-art late score fusion methods ([32] and [29]).

We assess the target identification performance in terms of False Alarm probability (FA), the probability that an impostor identity is accepted, and Miss-Detection probability (MD), the probability that a correct identity is rejected, as defined

Fig. 3: Sample images from the experimental dataset. (a) Targets from the *QM-Seminars* dataset. (c) Images from the *QM-GoPro* dataset.

by the acceptance/rejection threshold τ. The performance is compared using the Equal Error Rate (EER) computed with the Bosaris' toolkit [15] and obtained considering the value of τ for which the false alarm rate equals the miss detection rate (i.e. EER=FA=MD).

A. Datasets

We use two real challenging audio-visual databases, namely the *QM-Seminars* and the *QM-GoPro* datasets (Figure 3). The audio streams are at 48kHz, 16 bits and the video resolution is 1920x1080, at 25 frames per second. Each recording is split in consecutive 5-second-long segments (skipping the first and last 10 seconds). Each segment consists of $N_I = 125$ images and 240000 audio samples[2].

The *QM-Seminars* dataset consists of 16 participants giving the same 1-minute talk three times. The talks are recorded using a JVC GY-HM150E High Definition Camcorder and a Sennheiser ew 100-ENG G3 E-Band Wireless System as lapel microphone. The presenters move freely and generate considerable pose and appearance changes. Moreover, in some sequences significant illumination changes occur (see Figure 3a).

The *QM-GoPro* dataset captures interactions of 13 participants speaking for 1 minute to a person wearing a chest-mounted GoPro camera. Speakers are up to a few metres from the microphones (distant-talking task) that are partially

[2]To facilitate comparisons these datasets are available at: http://www.eecs.qmul.ac.uk/~andrea/adaptation.html

Fig. 4: Target appearance variations across different sequences of the same dataset. (Top) Target #14 from the *QM-Seminars* dataset. (Bottom) A target from the *QM-GoPro* dataset.

covered by the plastic shield of the camera. The dataset includes four conditions: indoors (C1); indoors wearing different clothes (C2); outdoors in a quiet location and wearing the same clothes as in C1 (C3); outdoors in a noisy location near a road with traffic and wearing the same clothes as in C2 (C4). The appearance of the target changes continuously within the same session and in some cases the face of the target is not even visible. The audio in C3 and C4 is affected by strong background noise and a considerable mismatch from the clean material used in the training of the feature extractor. Figure 4 (bottom) shows an example of the variety of environmental conditions and target appearances in the dataset, making the identification task particularly challenging.

B. Experimental setup

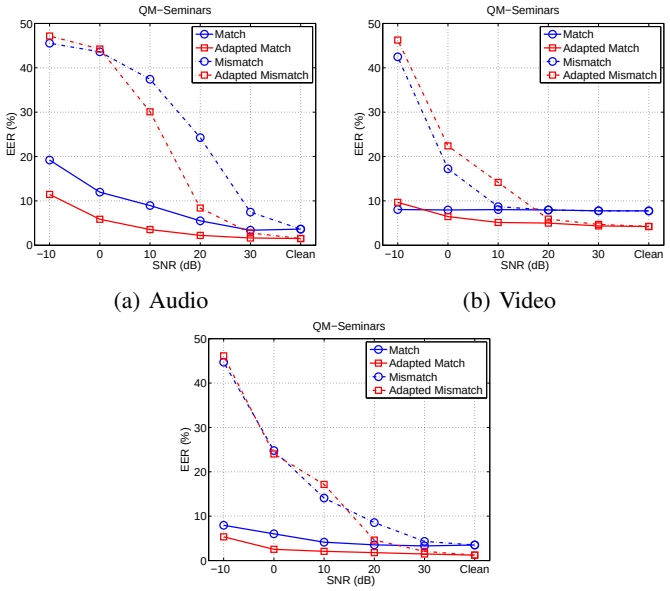
The bounding boxes of the targets are estimated with the ACF image-based detector available in the Piotr's toolbox [27], trained on the Caltech [26] and INRIA [20] pedestrian datasets. We extract the upper body considering a fixed size of 180x420 pixels, split in nine non-overlapping sub-images forming a 3-by-3 grid. A feature vector is created from the 5-bin RGB histograms of each sub-image, except for the top-left and top-right sub-images.

In order to use all the targets in the two datasets for testing, the TV feature extractor was trained on the out-of-domain data of the clean Italian APASCI dataset [8]. The dimension of the final i-Vectors is 400.

In the final fusion of the modalities, we use a constant value for the relevance factor: $r = 1$. Finally, the T-norm [10] is applied removing the average score of all targets:

$$\overline{c}^i(s,t) = c^i(s,t) - \frac{1}{S}\sum_{s=1}^{S} c^i(s,t). \tag{31}$$

In the *QM-Seminars* dataset, the target models are acquired using the first 6 segments of the first talk. All the other segments are used for testing, thus resulting in 476 trials (7140 impostor trials). We contaminated the recordings to create different degrees of mismatch by adding white noise to the



(a) Audio (b) Video

(c) Late fusion

Fig. 5: EER results on the *QM-Seminars* dataset for the two mono-modal systems ((a): audio; (b): video) and the multi-modal system (c) when models are blindly adapted, compared with the original models.

audio and video streams, varying the Signal-to-Noise Ratio (SNR) from -10 dB to 30 dB. In the *mismatched* case, models derived from clean signals are used in all the SNR conditions. In *matched* experiments, models are trained on the same noisy conditions as the test signals.

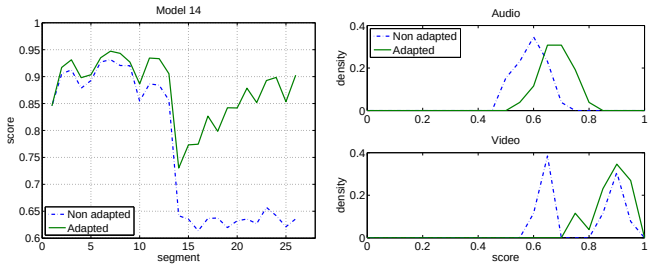
In the *QM-GoPro* dataset, target models are acquired using the first 3 segments of each condition. All the other segments are used for testing, thus resulting in an average of 86 trials (1032 impostors). In mismatched conditions, models from C1 are used to recognise the target identities in the other 3 scenarios. Instead of considering nine upper-body regions, a single RGB histogram vector is extracted in this case. This solution is more appropriate for the *QM-GoPro* dataset because of the varying distance of the target from the camera.

For a fair evaluation of the adaptation, targets appear in random order and remain in the scene for at least 25 seconds. Results are averaged over 20 repetitions, randomly varying the target order.

C. QM-Seminars dataset: results

Figures 5a and 5b show the performance of the two mono-modal systems with and without the proposed model adaptation in matched and mismatched conditions. Without adaptation, audio outperforms video in high SNR ($>$ 10 dB in matched conditions and $>$ 30 dB in mismatch), suggesting that the i-Vector framework better discriminates the subjects of this dataset than the colour histograms. Conversely, video is robust against noise: in matched conditions the performance is independent of noise, while in mismatch the performance starts deteriorating at 0 dB.

When the unsupervised multi-modal model adaptation is employed, noticeable performance improvements are obtained



(a) Video scores (b) Score distributions

Fig. 6: Scores for model #14 when target #14 is in the scene. (a) The video classification score suddenly drops after segment 13 (end of talk-2) using the non adapted model, while the adapted model converges to the new appearance. (b) Distributions of audio and video scores, with and without adaptation.

on the audio modality, in all conditions except the very low SNR cases in model mismatch. A similar improvement is obtained for the video system. However, as the SNR degrades, the adapted video models perform worse than the non-adapted ones. Already at 10 dB, the original models outperform the adapted ones in the presence of mismatch. This behaviour is related to how audio models adapt to mismatch conditions and will be discussed in Section VI-E.

When the SNR is above 10 dB, the adaptation of the video model to changes in target appearance is very effective. As an example, the upper part of Figure 4 shows 4 sample images of target #14. A change in illumination occurs between enrolment and talk-3. In addition to this, the target is often partially mis-detected in talk-2. Figure 6 compares the video scores obtained on each segment of target #14 with the original model and with the adapted one. Note in Figure 6a how the change in illumination leads to considerable low scores in test segments of talk-3 (second half). The adapted model alleviates the impact of the appearance change by moving segment-by-segment towards the new model. Figure 6b shows the distribution of the audio and video scores of target #14 with and without adaptation. Note that besides the video modality (bottom panel), the audio processing scores (top panel) also benefit from the adaptation.

Figure 5c shows the performance of the full multi-modal system with adaptation and late fusion, compared with the performance of the late fusion on the non-adapted models. The adaptation improves the performance in matched conditions (all SNRs) and in mismatched conditions when the SNR is above 10 dB. In mismatched conditions, the deterioration of the video models (Figure 5b) results in no gain when using adapted models in the late score fusion. Finally, note that the adopted score fusion is not effective in the presence of high mismatch as the multi-modal system performs worse than the video system for SNRs lower than 10 dB (compare Figure 5c and Figure 5b). This issue will be addressed in section VI-E.

Finally, Figure 7 shows the joint distributions of the audio and video scores for three good SNR conditions (Clean, 30 dB and 20 dB) using matched models, with and without adaptation. Note how adaptation leads to a better discrim-

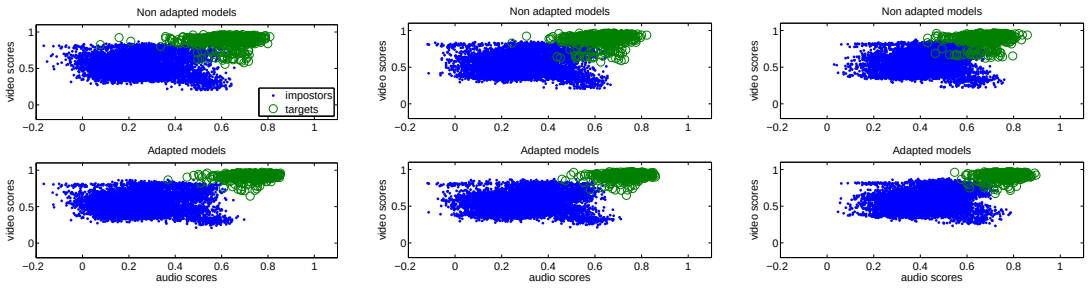


Fig. 7: Joint distribution of the audio and video scores on the *QM-Seminars* dataset in matched conditions, with and without model adaptation. Note how the model adaptation generates a better separation between the target (green circles) and the impostor scores (blue dots).

ination between the target and impostors scores, increasing the separation between the two clusters of score vectors. This fact implies that model adaptation would also be beneficial if coupled with other late-fusion strategies.

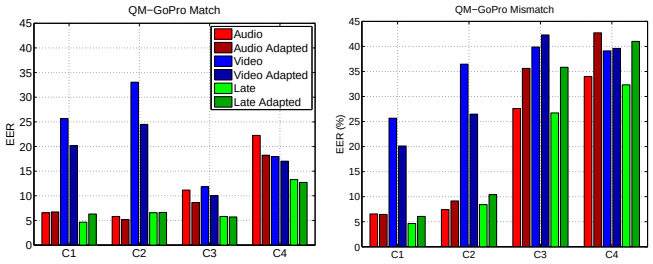
D. QM-GoPro dataset: results

Figure 8a shows the EER results with and without the proposed model adaptation scheme. As expected, audio behaves particularly well in C1 and C2 as there are no particular challenges and the acoustic propagation is constant. Performance deteriorates in C3 and C4 due to environmental noise.

The indoor scenarios (C1 and C2) are characterised by noticeable and continuous changes in the illumination and in the target positions, thus leading to poor performance of the the video modality. The experimental conditions are slightly better in C3 and C4 as the targets behave in a similar way to the *QM-Seminars* dataset. Nevertheless, the performance is slightly worse probably due to the different quality of the video sensors and the continuous movements of the person wearing the camera. Finally, except for C2, late fusion brings a considerable improvement.

The impact of model adaptation varies considerably. In C1 and C2 the two modalities perform very differently and the quality of the video signal is low. Model adaptation improves the video performance at the cost of a minor deterioration of the audio performance. Since the two systems tend to adapt towards a common agreement on the models, late fusion does not help and slightly deteriorates the results. This problem can be addressed by properly setting the relevance factor, as explained in Section VI-E. Conversely, in C3 and C4 the two systems perform similarly and the adaptation improves in particular the audio models, which still suffer from the mismatch in the feature extraction. For the reasons discussed above fusion only marginally improves over the single modalities and over the fusion of the non-adapted models.

Figure 8b shows the recognition performance in the mismatch case, when models are trained in condition C1 only. The performance for the video modality drops between the indoor scenarios C1 and C2 as the colour appearance of targets changes. A further performance degradation is observed in C3 and C4, when natural light replaces artificial indoor lighting. The audio modality has a minor deterioration between C1



(a) Matched (b) Mismatched

Fig. 8: EER results on the *QM-GoPro* dataset in matched and mismatched conditions for the four scenarios. In mismatch, models of C1 are used in the other conditions.

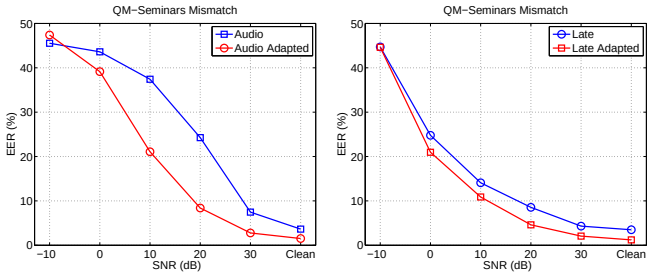
and C2. Conversely (and as expected) background noise is detrimental and considerably increases the recognition error in C3 and C4. Also in this case fusion improves the performance.

The model adaptation in the mismatch scenario is not very effective: in C1 and C2 the video models are improved but a deterioration in the audio performance and in the fusion is observed because the performance of the video modality is poor. The mismatch in C3 and C4 is too large and models cannot adapt successfully.

E. Discussion on the relevance factor

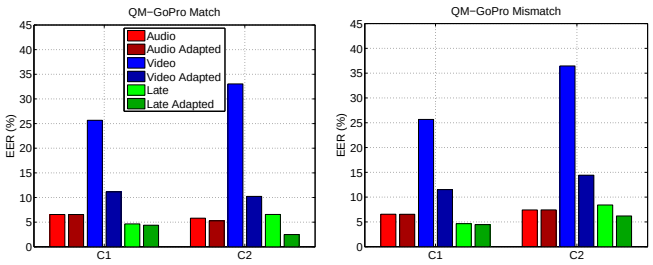
The adaptation fails when at least one of the two modalities has *very* poor performance, for example in scenarios C3 and C4 (see Figure 8b). However, when only one of the two models is *very* poor, its correction is possible. To increase robustness in highly mismatched conditions, an adaptive relevance factor could be used to control the amount of adaptation from a given observation vector.

The performance deterioration in the *QM-Seminars* data (Figures 5b and 5c) occurs because of a high mismatch in the audio models, as confirmed by the very poor performance of the audio modality. As a consequence, the audio model of the first target appearing adapts to the noise (rather than to the new speaker model) and dominates all the other mismatched models in subsequent segments. This leads to one of the audio models collecting all the new observations, which are all assigned to the related video model. As a result, while the



(a) Video (b) Late

Fig. 9: EER results on the *QM-Seminars* dataset in mismatched conditions for the audio system and the late fusion when the video models are not adapted for SNRs below 20 dB. Note that the adapted late fusion achieves the same performance as the non-adapted video for low SNRs.

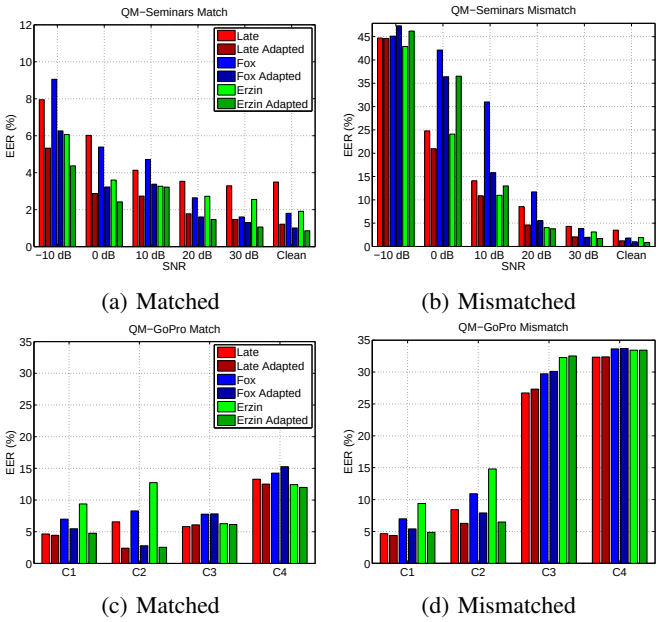


(a) Matched (b) Mismatched

Fig. 10: EER results on the *QM-GoPro* dataset in matched and mismatched conditions for scenarios C1 and C2. Different relevance factors are used according to the models accuracy: no adaptation for the audio modality and a relevance factor of 0.2 for video model adaptation.

performance of the audio modality still improves slightly, the video models deteriorate completely, thus affecting the late fusion. This problem can be partially alleviated by modifying the relevance factor so that the video models are not adapted. Figure 9 reports the EER of the audio system and of the late fusion in mismatched conditions when the video models are not adapted if the SNR is below 20 dB. Note that the performance of both audio and multi-modal systems are noticeably improved.

A similar problem occurs in the *QM-GoPro* dataset in scenarios C1 and C2. In this case the video models are extremely poor and negatively affect the adaptation of the audio models, thus deteriorating the performance of the audio system and of the late score fusion. Similarly to what observed above, this problem can be tackled by adopting suitable relevance factors that limit the adaptation of the accurate models and speed up the update of the erroneous models. Figure 10 shows the performance in C1 and C2 in both matched and mismatched conditions when the relevance factor of the audio is very high (i.e. very marginal adaptation of the models) and the relevance factor of the video is set to 0.2. The performance of the video model is considerably improved without affecting the audio modality. Note how the video appearance mismatch between C1 and C2 is almost completely compensated. Finally, a small improvement is also obtained on the final score fusion.



(a) Matched (b) Mismatched

(c) Matched (d) Mismatched

Fig. 11: EER results on the *QM-Seminars* (top) and *QM-GoPro* (bottom) datasets in matched and mismatched conditions using three late fusion approaches with and without model adaptation.

F. Comparative analysis

To complete the experimental analysis, we combine the proposed model adaptation with two state-of-the-art solutions based on the weighted score combinations presented by Fox [32] and Erzin [29]. We consider the final end-to-end system with the appropriate relevance factor. Figure 11(top) shows the results in matched and mismatched conditions for the *QM-Seminars* dataset. In the matched scenario the three fusion methods perform similarly. Importantly, they all benefit from the adaptation, leading to very similar EER. In the mismatched case, the proposed late fusion is superior to the one by Fox (except for the "Clean" case) and both benefit substantially from the adaptation. Conversely, the fusion by Erzin outperforms the other two methods but it does not improve with the adapted models for low SNRs (0 and -10 dB). This happens because Erzin uses as measure of confidence the absolute value of the score, which fails when the adapted audio models have larger scores, on average, due to noise. Finally, the results of the proposed system with adaptation are comparable with Erzin's.

Figure 11(bottom) shows the results of the comparative analysis on the *QM-GoPro* dataset. The proposed late fusion outperforms the state-of-the-art methods with and without adapted models. Moreover, all the late fusion approaches benefit from adaptation (excluding C3 and C4 in mismatched conditions, where adaptation is not possible for the reasons reported above) showing that the proposed approach is beneficial in a variety of multi-modal target (re-)identification pipelines.

VII. Conclusion

We proposed a generic framework to continuously adapt audio and video models for multi-modal person recognition. The proposed method is unsupervised and exploits the complementarity of audio and video information. The proposed adaptation improves the performance of models trained on out-of-domain data and tackles changes in target appearance and in the background. Importantly, our solution is independent of the specific audio and video processing implementation and can be combined with different front-ends. Moreover, with our approach it is straightforward to add other visual features, for example produced via deep learning [3], [5], [41] or based on biometrics [6], [29], [44].

Future work includes the automatic detection of unseen targets and the related on-the-fly model acquisition as well as the development of more sophisticated training strategies, such as Probabilistic Linear Discriminant Analysis (PLDA) [55], [57]. Other open research issues include the definition of an adaptive relevance factor to be linked to the confidence measure of the individual classifiers and the investigation of a more robust algorithm for late score integration, possibly based on a further classification stage.

References

[1] http://trecvid.nist.gov/, Last access on April 20th 2016.

[2] A. Abad, "The L2F language recognition system for NIST LRE 2011," in *The NIST Language Recognition Evaluation Workshop*, 2011.

[3] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, USA, 2015.

[4] M. R. Alam, M. Bennamoun, R. Togneri, and F. Sohel, "A confidence-based late fusion framework for audio-visual biometric identification," *Pattern Recognition Letters*, vol. 52, 2015.

[5] ——, *Image and Video Technology: 7th Pacific-Rim Symposium, Revised Selected Papers*. Springer International Publishing, 2016, ch. Deep Boltzmann Machines for i-Vector Based Audio-Visual Person Identification.

[6] A. Albiol, L. Torres, and E. Delp, "Fully automatic face recognition system using a combined audio-visual approach," *IEE Proceedings of Vision, Image and Signal Processing*, vol. 152, no. 3, Jun. 2005.

[7] P. S. Aleksic and A. K. Katsaggelos, "Audio-visual biometrics," *Proc. of the IEEE*, vol. 94, no. 11, 2006.

[8] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo, "Speaker independent continuous speech recognition using an acoustic-phonetic italian corpus," in *Proc. of the Int. Conf. on Spoken Language Processing*, Yokohama, Japan, 1994.

[9] J. J. Atick, P. M. Griffin, and A. N. Redlich, "Faceit: face recognition from static and live video for law enforcement," in *Proc. of SPIE*, vol. 2932, 1997.

[10] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, 2000.

[11] C. Barras, S. Meignier, and J.-L. Gauvain, "Unsupervised online adaptation for speaker verification over the telephone," in *Proc. of Speaker Odyssey*, Toledo, Spain, 2004.

[12] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing*, vol. 32, no. 4, 2014.

[13] S. Bengio, C. Marcel, S. Marcel, and J. Mariethoz, "Confidence measures for multimodal identity verification," *Information Fusion*, 2002.

[14] S. Bickel and T. Scheffer, "Estimation of mixture models using co-EM," in *Proc. of the ICML Workshop on Learning with Multiple Views*, Bonn, Germany, 2005.

[15] N. Brummer and E. de Villiers, "The BOSARIS Toolkit User Guide: Theory, Algorithms and Code for Binary Classifier Score Processing," Tech. Rep., 2011.

[16] D. Burnham, D. Estival, S. Fazio, F. Cox, R. Dale, J. Viethen, S. Cassidy, R. Togneri, Y. Kinoshita, J. Arciuli, M. Onslow, T. Lewis, A. Butcher, and J. Hajek, "Building an audio-visual corpus of australian english: large corpus collection with an economical portable and replicable black box," in *Proc. of Interspeech*, Florence, Italy, 2011.

[17] J. Carletta, "Announcing the AMI meeting corpus," in *The ELRA Newsletter*, January-March 2006, vol. 11, no. 1.

[18] U. V. Chaudhari, G. N. Ramaswamy, G. Potamianos, and C. Neti, "Audio-visual speaker recognition using time-varying stream reliability prediction," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 5, Hong Kong, 2003.

[19] C. M. Christoudias, K. Saenko, L.-P. Morency, and T. Darrell, "Co-adaptation of audio-visual speech and gesture classifiers," in *Proc. of the Int. Conf. on Multimodal Interfaces*, 2006.

[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, San Diego, USA, 2005.

[21] N. Dehak, P. 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, May 2011.

[22] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Cosine similarity scoring without score normalization techniques," in *Proc. of Speaker Odyssey*, Brno, Czech Republic, 2010.

[23] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proc. of Interspeech*, Brighton, United Kingdom, 2009.

[24] 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*, vol. 39, no. 1, 1977.

[25] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, Aug. 2014.

[26] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, Apr. 2012.

[27] P. Dollár, "Piotr's Computer Vision Matlab Toolbox (PMT)." [Online]. Available: http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html

[28] H. Ekenel, M. Fischer, Q. Jin, and R. Stiefelhagen, "Multi-modal person identification in a smart environment," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Minneapolis, USA, 2007.

[29] E. Erzin, Y. Yemez, and A. Tekalp, "Multimodal speaker identification using an adaptive classifier cascade based on modality reliability," *IEEE Transactions on Multimedia*, vol. 7, no. 5, Oct. 2005.

[30] E. Erzin, Y. Yemez, and A. Murat Tekalp, "Joint audio-video processing for robust biometric speaker identification in car," in *DSP for In-Vehicle and Mobile Systems*, H. Abut, J. Hansen, and K. Takeda, Eds. Springer US, 2005.

[31] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, USA, 2010.

[32] N. Fox, R. Gross, J. Cohn, and R. Reilly, "Robust biometric person identification using automatic classifier fusion of speech, mouth, and face experts," *IEEE Transactions on Multimedia*, vol. 9, no. 4, Jun. 2007.

[33] J. Gauvain and C.-H. Lee, "Maximum A Posteriori estimation for multivariate gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, Apr. 1994.

[34] V. Hautamaki, T. Kinnunen, I. Karkkainen, J. Saastamoinen, M. Tuononen, and P. Franti, "Maximum A Posteriori adaptation of the centroid model for speaker verification," *Signal Processing Letters*, vol. 15, 2008.

[35] T. J. Hazen, K. Saenko, C.-H. La, and J. R. Glass, "A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments," in *Proc. of the Int. Conf. on Multimodal Interfaces*, New York, USA, 2004.

[36] T. J. Hazen, E. Weinstein, B. Heisele, A. Park, and J. Ming, *Face Biometrics for Personal Identification: Multi-Sensory Multi-Modal Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, ch. Multimodal Face and Speaker Identification for Mobile Devices.

[37] L. Heck and N. Mirghafori, "On-line unsupervised adaptation in speaker verification: Confidence-based updates and improved parameter estimation," in *ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*, 2001.

[38] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, 2010.

[39] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, Mar. 1998.

[40] A. Levin, P. Viola, and Y. Freund, "Unsupervised improvement of visual detectors using co-training," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Washington, USA, 2003.

[41] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Columbus, USA, 2014.

[42] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Portland, USA, 2013.

[43] S. Lucey, T. Chen, S. Sridharan, and V. Chandran, "Integration strategies for audio-visual speech processing: applied to text-dependent speaker recognition," *IEEE Transactions on Multimedia*, vol. 7, no. 3, Jun. 2005.

[44] J. Luque, R. Morros, A. Garde, J. Anguita, M. Farrus, D. Macho, F. Marqués, C. Martínez, V. Vilaplana, and J. Hernando, "Audio, video and multimodal person identification in a smart room," in *Multimodal Technologies for Perception of Humans: First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2006*, R. Stiefelhagen and J. Garofolo, Eds., 2007.

[45] A. Ma, J. Li, P. Yuen, and P. Li, "Cross-domain person reidentification using domain adaptation ranking SVMs," *IEEE Transactions on Image Processing*, vol. 24, no. 5, May 2015.

[46] R. Mazzon, S. F. Tahir, and A. Cavallaro, "Person re-identification in crowd," *Pattern Recognition Letters*, vol. 33, no. 14, Oct. 2012.

[47] C. McCool, S. Marcel, A. Hadid, M. Pietikainen, P. Matejka, J. Cernocky, N. Poh, J. Kittler, A. Larcher, C. Levy, D. Matrouf, J.-F. Bonastre, P. Tresadern, and T. Cootes, "Bi-modal person recognition on a mobile phone: using mobile phone data," in *Proc. of ICME Workshop on Hot Topics in Mobile Multimedia*, Melbourne, Australia, 2012.

[48] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard, "Modeling human interaction in meetings," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Hong Kong, 2003.

[49] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. of the Int. Conf. on Audio and Video-based Biometric Person Authentication*, Washington, USA, 1999.

[50] N. Moreau, D. Mostefa, R. Stiefelhagen, S. Burger, and C. K., "Data collection for the CHIL CLEAR 2007 evaluation campaign," in *Proc. of the Int. Conf. on Language Resources and Evaluation*, Marrakech, Morocco, 2008.

[51] W. Nam, P. Dollár, and J. H. Han, "Local decorrelation for improved detection," *Neural Information Processing Systems*, 2014.

[52] J. R. Norris, *Markov chains*, ser. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, 1998.

[53] A. Noulas, G. Englebienne, and B. Krose, "Multimodal speaker diarization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, Jan. 2012.

[54] A. Preti and J.-F. Bonastre, "Unsupervised model adaptation for speaker verification," in *Proc. of Interspeech*, Pittsburgh, USA, 2006.

[55] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Rio de Janeiro, Brazil, 2007.

[56] B. Prosser, S. Gong, and T. Xiang, "Multi-camera matching using bi-directional cumulative brightness transfer functions," in *Proc. of the British Machine Vision Conf.* Leeds, United Kingdom: BMVA Press, 2008.

[57] P. Rajan, A. Afanasyev, V. Hautamki, and T. Kinnunen, "From single to multiple enrollment i-Vectors: Practical PLDA scoring variants for speaker verification," *Digital Signal Processing*, vol. 31, 2014.

[58] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, 2000.

[59] C. Sanderson, S. Bengio, H. Bourlard, J. Mariethoz, R. Collobert, M. BenZeghiba, F. Cardinaux, and S. Marcel, "Speech face based biometric authentication at IDIAP," in *Proc. of the Int. Conf. on Multimedia and Expo*, Baltimore, USA, 2003.

[60] C. Sanderson and K. K. Paliwal, "Noise compensation in a person verification system using face and multiple speech features," *Pattern Recognition*, vol. 36, no. 2, 2003.

[61] ——, "Identity verification using speech and face information," in *Digital Signal Processing*, vol. 14, no. 5, 2004.

[62] R. Seymour, M. Ji, and D. Stewart, "A new posterior based audio-visual integration method for robust speech recognition," in *Proc. of Interspeech*, Lisbon, Portugal, 2005.

[63] S. Shivappa, M. Trivedi, and B. Rao, "Audiovisual information fusion in human computer interfaces and intelligent environments: A survey," *Proc. of the IEEE*, vol. 98, no. 10, Oct. 2010.

[64] P. Smaragdis and M. Casey, "Audio/visual independent components," in *Proc. of the Int. Symposium on Independent Component Analysis and Blind Source Separation*, 2003.

[65] R. Stiefelhagen, K. Bernardin, R. Bowers, R. Rose, M. Michel, and J. Garofolo, "The CLEAR 2007 Evaluation Multimodal Technologies for Perception of Humans," in *Multimodal Technologies for Perception of Humans*, ser. Lecture Notes in Computer Science, R. Stiefelhagen, R. Bowers, and J. Fiscus, Eds. Berlin, Heidelberg: Springer Berlin / Heidelberg, 2008, ch. 1.

[66] C. Taskiran, J.-Y. Chen, A. Albiol, L. Torres, C. Bouman, and E. Delp, "Vibe: a compressed video database structured for active browsing and search," *IEEE Transactions on Multimedia*, vol. 6, no. 1, Feb. 2004.

[67] R. Vezzani, D. Baltieri, and R. Cucchiara, "People reidentification in surveillance and forensics: A survey," *ACM Computing Surveys*, vol. 46, no. 2, Dec. 2013.

[68] Y. Yan, E. Ricci, G. Liu, and N. Sebe, "Egocentric daily activity recognition via multitask clustering," *IEEE Transactions on Image Processing*, vol. 24, no. 10, 2015.

[69] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, "Modeling individual and group actions in meetings with layered HMMs," *IEEE Transactions on Multimedia*, vol. 8, no. 3, Jun. 2006.

[70] H. Zhang, V. Patel, S. Shekhar, and R. Chellappa, "Domain adaptive sparse representation-based classification," in *Proc. of the IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition*, Ljublijana, Slovenia, 2015.

Alessio Brutti is a tenured researcher at Fondazione Bruno Kessler, Trento, Italy. After graduating in Telecommunication engineering at the University of Padova, Padova, Italy, in 2001, in 2003 he joined the Center for Information and Communication Technologies of FBK. In 2006 he completed his Ph.D. in Computer Science at the University of Trento, Trento, Italy. His main research interests focus on multichannel digital audio processing and include: localization and tracking of acoustic sources, speaker verification and recognition, acoustic scene analysis and speech enhancement for speech recognition in adverse conditions. He is also active in multi-modal signal processing, in particular for audio-video person tracking and identification.

Andrea Cavallaro is Professor of Multimedia Signal Processing and Director of the Centre for Intelligent Sensing at Queen Mary University of London, UK. He received his Ph.D. in Electrical Engineering from the Swiss Federal Institute of Technology (EPFL), Lausanne, in 2002. He was a Research Fellow with British Telecommunications (BT) in 2004/2005 and was awarded the Royal Academy of Engineering teaching Prize in 2007; three student paper awards on target tracking and perceptually sensitive coding at IEEE ICASSP in 2005, 2007 and 2009; and the best paper award at IEEE AVSS 2009. Prof. Cavallaro is Associate Editor for the IEEE Transactions on Circuits and Systems for Video Technology and member of the editorial board of IEEE Multimedia. He is a past Area Editor for IEEE Signal Processing Magazine and a past Associate Editor for the IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, IEEE Transactions on Signal Processing, and IEEE Signal Processing Magazine. He has published over 160 journal and conference papers, one monograph on Video tracking (2011,Wiley) and three edited books: Multi-camera networks (2009, Elsevier); Analysis, retrieval and delivery of multimedia content (2012, Springer); and Intelligent multimedia surveillance (2013, Springer).