Unsupervised cross-modal deep-model adaptation for audio-visual re-identification with wearable cameras

Alessio Brutti
Center for Information Technology
Fondazione Bruno Kessler, Trento, Italy
brutti@fbk.eu

Andrea Cavallaro
Centre for Intelligent Sensing
Queen Mary University London, UK
a.cavallaro@qmul.ac.uk

Abstract

Model adaptation is important for the analysis of audio-visual data from body worn cameras in order to cope with rapidly changing scene conditions, varying object appearance and limited training data. In this paper, we propose a new approach for the on-line and unsupervised adaptation of deep-learning models for audio-visual target re-identification. Specifically, we adapt each mono-modal model using the unsupervised labelling provided by the other modality. To limit the detrimental effects of erroneous labels, we use a regularisation term based on the Kullback-Leibler divergence between the initial model and the one being adapted. The proposed adaptation strategy complements common audio-visual late fusion approaches and is beneficial also when one modality is no longer reliable. We show the contribution of the proposed strategy in improving the overall re-identification performance on a challenging public dataset captured with body worn cameras.

1. Introduction

Audio-visual processing is a key step in applications such as surveillance [40], meeting analysis [38], biometrics [3] and audio-visual speech recognition [1, 39]. Independently of the specific task, most solutions rely on late score fusion or, to a lesser extent, on early feature concatenation [29, 36].

For the person re-identification task, a variety of audio-visual fusion methods with sophisticated combination strategies for the data from the two sensing modalities have been proposed (see [9] for an overview). However, these solutions 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 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 generate reliable person models to be used for re-identifying the target in future interactions. Moreover, other important issues unaddressed by state-of-the-art solutions include frequent variations of target appearance and environmental conditions as well as the intermittent availability of information when a target moves outside the field of view or becomes silent.

Recently, Deep Neural Networks (DNNs) have led to substantial performance improvements under varying operational conditions in speech recognition [5], speaker verification [34] and visual object tracking [7]. Nevertheless, model adaptation to varying operational conditions or to different application domains is still an unsolved problem. In particular, when the adaptation is unsupervised, care is needed to prevent erroneous labels from negatively affecting the quality of the model [30, 33].

The audio-visual target re-identification problem can be addressed via information fusion and by cross-labelling for unsupervised adaptation [9], leveraging the complementarity of the information available from audio and visual streams [36]. Cross-modal unsupervised labelling is important when the amount of training data is insufficient. However, labelling errors could overshadow the benefits of the learning capability of a DNN.

In this paper, we extend the multi-modal adaptation concepts of [9] with DNNs [23] and use a Multi-Layer Perceptron (MLP) in the target classification stage. Moreover, we introduce a regularisation term based on the Kullback-Leibler Divergence (KLD) to avoid model divergence when unsupervised labelling is unreliable, as done in unsupervised DNN adaptation for speech recognition [18, 42]. In particular, we consider two mono-modal systems followed by a late score fusion stage. The proposed approach is more attractive than early feature concatenation because each modality is processed independently, thus allowing the design of each mono-modal system to match the

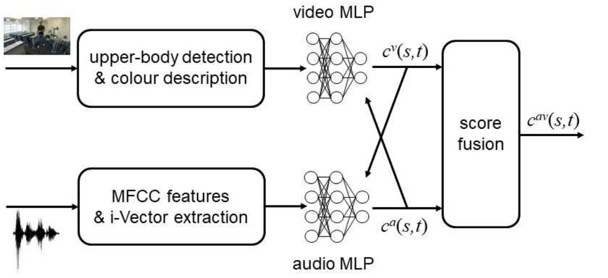

video MLP
upper-body detection & colour description
$c^v(s,t)$
score fusion
$c^{av}(s,t)$
MFCC features & i-Vector extraction
$c^a(s,t)$
audio MLP

Figure 1: Block diagram of the proposed audio-visual target re-identification system.

properties of the observations. After feature extraction, two independent MLPs identify the target using audio and visual information, respectively. The outputs of each MLP are used to label the observations of the other modality in the adaptation stage. The block diagram of the proposed audio-visual target re-identification scheme is shown in Figure 1.

This paper is organised as follows. Sec. 2 reviews related works on unsupervised model adaptation and highlights the novel contribution of this paper. Sec. 3 introduces the audio-visual target re-identification problem and presents our general framework. Sec. 4 describes the proposed model adaptation based on cross-modality labelling. Sec. 5 describes the dataset, provides details about our implementation and discusses the experimental results. Finally, Sec. 6 concludes the paper with final remarks and plans for future work.

2. Related work

Robust unsupervised model adaptation approaches exist that use mono-modal information only for (visual) person re-identification [30, 43] and for speaker verification [6, 22, 33]. In [20] domain adaptation is used for off-line speaker verification on an unlabelled training set.

When multiple *views* of the same data are available, co-EM [8] can be employed to iteratively adapt each model by exchanging labels and by minimising the disagreement between modalities. Co-EM is referred to as co-adaptation in speech recognition [11] and as co-training in multi-camera traffic analysis [27]. Co-adaptation is used for audio-visual speech recognition and gesture recognition [11]. The audio and visual models are jointly adapted using unseen data of the application domain which are labelled by maximizing the agreement of the audio and visual modalities.

In our problem an adaptation set is not available and we instead adapt models on-line as soon as a new observation is available. This constraint does not allow the use of domain adaptation or transfer learning techniques.

Adaptation in DNNs requires care to avoid over-fitting the unlabelled adaptation data, using for example regularisation. In Automatic Speech Recognition (ASR) applications, this is achieved by retraining only subsets of the network weights, namely those of the input layer [24, 35] or of the output layer [41]. Other methods apply linear transformations to input, output or hidden DNN layers [2, 21, 28, 32, 37]. The use of dropout for regularisation has also been investigated [31]. Finally, the application of a momentum term to update the DNN weights, the introduction of regularisation terms, the use of small values for the learning rate as well as of an early stopping criterion can regularise a given original model (Conservative Training (CT) [42]).

The above-mentioned methods address the unsupervised adaptation of mono-modal models. In this paper, we apply the KLD-regularisation used in [18] to cross-modal model adaptation.

3. Multi-modal re-identification

3.1. Target re-identification

Let $\mathcal{S}$ be a set of S enrolled targets and $p(y_t = s|\mathbf{x}_t)$ the probability that $\mathbf{x}_t$, the t-th[1] observation, is generated by a target with identity $y_t = s$, where $t = 1, \ldots, T$ and T is the number of observations. The target identity $\hat{y}_t$ is estimated as:

$$\hat{y}_t = \arg\max_{s\in\mathcal{S}} p(y_t = s|\mathbf{x}_t) \tag{1}$$

For each modality the posterior is computed using a deep MLP fed with the features described in Sec. 3.2 and trained using the Negative Log-Likelihood (NLL), the typical multi-class regression cost:

$$\begin{aligned}\mathcal{L}(\Theta|\mathbf{x}) = -\sum_{t=1}^{T}\sum_{s\in\mathcal{S}} \tilde{p}(y_t = s|\mathbf{x}_t)\log(p(y_t = s|\mathbf{x}_t)) \quad (2)\\ + \lambda\|w\|^2,\end{aligned}$$

where $\mathbf{x}$ is the set of input vectors for training, Θ is the set of network parameters (i.e. weights w and bias b), $\tilde{p}(y_t = s|\mathbf{x}_t)$ is the target probability distribution, and λ is the parameter that controls the L2 regularisation term, $\lambda\|w\|^2$. If we assume that $\tilde{p}(y_t = s|\mathbf{x}_t) = \delta(y_t - \tilde{y}_t)$, then Eq. 2 becomes:

$$\mathcal{L}(\Theta|\mathbf{x}) = -\sum_{t=1}^{T}\log(p(\tilde{y}_t|\mathbf{x}_t)) + \lambda\|w\|^2, \tag{3}$$

where $\tilde{y}_t$ is the label associated to observation $\mathbf{x}_t$ (target class).

The score for target s is obtained for observation $\mathbf{x}_t^i$ for modality i by taking the MLP output:

$$c^i(s,t) = p(y_t = s|\mathbf{x}_t^i). \tag{4}$$

[1] We use t as a generic index of the observations or feature vectors. It could be time, frame index or utterance index, depending on the feature type.

3.2. Audio and visual features

To limit the computational complexity of the adaptation, we avoid end-to-end solutions operating directly on the RGB images or on the audio signals. Moreover, to ease cross-modal interaction we design similar pipelines for the audio and the visual sub-systems.

We extract the *visual* features from the upper body of the target. To estimate the bounding box of a target we use the Aggregate Channel Features (ACF) image-based detector of the Piotr's toolbox [15, 16], trained on the Caltech [17] and INRIA [12] pedestrian datasets. Then, we extract the upper body by simply considering a sub-image of fixed size (180x420 pixels), centred horizontally and whose upper side is 30 pixels below the upper side of the detected bounding box. From the estimated upper body we extract and concatenate three 21-bin RGB colour histograms to generate a 63-dimensional feature vector. The MLP network for classification has 5 layers of 2048, 1024, 1024, 512 and 512 neurons respectively. The activation function is $\tanh$. On top of the DNN we use a soft-max regression layer.

The *audio* features are based on the Total Variability (TV) paradigm [13, 14]. We generate 30-dimensional feature vectors by concatenating 15 Mel-Frequency Cepstral Coefficients (MFCC) with their first derivatives (energy is not used). We extract the coefficients from 20 ms windows, with 10 ms steps. Starting from the MFCCs and using a TV feature extractor trained on the out-of-domain clean Italian APASCI dataset [4], we extract 200-dimensional I-vectors [13, 14] for each utterance. The audio network has 2 layers with 2048 and 1024 neurons and a final soft-max regression layer. The activation is again $\tanh$. The audio MLP network is smaller than the video network because the training material is more limited: for each sentence one I-vector is available.

In both modalities, the L2 regularisation parameter λ is 10^{-4} and the learning rate is 0.1 with a 10% decrement at each epoch.

3.3. Multi-modal fusion

Several combination strategies could be used to fuse the target scores, $c^a(s,t)$ and $c^v(s,t)$, produced by the two DNNs.

The use of a further classification stage that learns to optimally combine scores, for example based on Recurrent Neural Networks (RNNs) [19], could be particularly efficient. However, this option cannot be used in our scheme as the on-line adaptation of the models would modify the distribution of the classification scores, thus requiring a further adaptation of the combination layers, which may be unachievable as it would require re-training them on a development set. We therefore combine the audio and visual scores with the sum rule [25]:

$$c^{av}(s,t) = \gamma_t c^a(s,t) + (1-\gamma_t) c^v(s,t), \tag{5}$$

where the weights, γ_t, are derived from the reciprocal of the variance of the classification scores, excluding the highest one [9]:

$$\gamma_t = \frac{\xi_t^a}{\xi_t^a + \xi_t^v}, \tag{6}$$

where

$$\xi_t^i = \frac{1}{\frac{1}{S-2}\sum_{s \neq s_{\text{ML}}^i} \left(c^i(s,t) - \mu_t^i\right)^2}, \tag{7}$$

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

4. Model adaptation

4.1. Exploiting multi-modality

When some supervision about the target class of the incoming feature vector is available during testing, DNN-based models can be adapted with further back-propagation iterations, typically in combination with small learning rates [42]. However, supervision on the test data is normally unavailable, thus making adaptation a considerably harder problem.

To address this issue, when different observations of the same event are available from different sensor types, the complementarity of multiple modalities can be exploited. We therefore use a cross-modal feedback to control the adaptation of each model given the current observation (see Fig. 1). In the specific implementation for this paper, the adaptation stage is a single training pass using a single observation with a small learning rate of 0.001, as a CT strategy [18, 42].

Let us assume that we are adapting the model parameters Θ^i of modality i. Eq. 3 can be modified to include labels estimated with the model i itself (*intra-modal adaptation*) or with another modality j (*cross-modal adaptation*). The cost function for cross-modal adaptation becomes:

$$\mathcal{L}(\Theta^i | \mathbf{x}^i) = -\sum_{t=1}^{T} \log(p(\hat{y}_t^j | \mathbf{x}_t^i)) + \lambda \|w\|^2, \tag{8}$$

where the class $\hat{y}_t^j$ estimated with modality j replaces the supervised labelling $\tilde{y}_t$.

Our goal is to use the t-th incoming observation feature vectors to adapt the temporal models Θ_t^i to the varying conditions. Therefore, Eq. 8 is reformulated by removing the summation over t:

$$\mathcal{L}(\Theta_t^i | \mathbf{x}_t^i, \hat{y}_t^j) = -\log(p(\hat{y}_t^j | \mathbf{x}_t^i)) + \lambda \|w\|^2. \tag{9}$$

The use of limited data for adaptation is a considerable challenge that could affect the models Θ_t^i in case of erroneous labels, even in presence of small learning rates. This problem will be addressed in the next section.

4.2. KLD regularisation

To avoid over-fitting or learning from erroneously labelled observations, we modify the training cost in Eq. 9 by introducing a further regularisation term based on the KLD between the model being adapted and the original one [42]. The idea is to force the network to preserve information about the original output distribution.

If $\overline{p}(y_t^i = s|\mathbf{x}_t^i)$ is the output of the original network for modality i, the new cost function becomes:

$$\tilde{\mathcal{L}}(\Theta_t^i|\mathbf{x}_t^i, \hat{y}_t^j) = (1-\rho)\mathcal{L}(\Theta_t^i|\mathbf{x}_t^i, \hat{y}_t^j) - \rho \sum_{s=1}^{S} \overline{p}(y_t^i = s|\mathbf{x}_t^i) \log(p(y_t^i = s|\mathbf{x}_t^i)), \quad (10)$$

where ρ controls the amount of adaptation. If ρ is small, the network is constrained towards the original output distribution and the observation is not used for adaptation. When ρ is large the network parameters are instead adapted.

Several options exist for tuning the adaptation parameter ρ. For example, ρ can be related to an estimation of the accuracy of the automatic labelling [18]. In our work, we relate the amount of regularisation to the posterior associated to the selected target, which can be considered as a measure of the reliability of the classification output.

We use different functions for video and for audio adaptation because the range of the output scores of the two networks is considerably different. While the range of the video scores is 0-1, the range of the audio scores is in general much lower. This could be related to the smaller size of the audio network or the larger dimension of the audio feature vectors, which make the audio MLP less discriminative.

For video adaptation, we empirically derive ρ from the audio network output as follows:

$$\rho^v = \frac{1}{[1 + \exp(-(c^a - 0.3) \cdot 30)]}, \quad (11)$$

whereas for audio adaptation we use:

$$\rho^a = c^v. \quad (12)$$

5. Experimental analysis

We compare the classification accuracy of the proposed cross-modal model adaptation ("cross") with the baseline systems without adaptation ("base") and with intra-modal adaptation ("intra"). For the proposed method, we also evaluate its performance with and without KLD regularisation.

5.1. Data and setup

We use a challenging audio-visual database, the *QM-GoPro* dataset, which captures interactions of 13 participants speaking for 1 minute to a person wearing a chest-mounted GoPro camera[2].

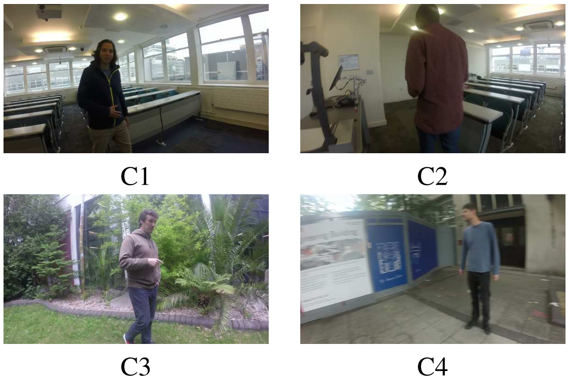

C1
C2
C3
C4

Figure 2: Sample frames for the four capturing conditions in the *QM-GoPro* dataset. Conditions C1 and C2 are challenging for video processing because illumination conditions change considerably as the speaker moves, e.g. towards or away from the windows. In addition, the furniture often partially occludes the body thus compromising the performance of the person detector. The distance of the speaker from the camera varies considerably (in some cases the target is so close that only a body part is visible), thus affecting the quality of the upper-body extraction, which uses a fixed bounding box size. Conditions C3 and C4 are simpler as the illumination conditions are constant and the target moves within a short range from the camera.

The dataset includes four recording conditions: indoors, in a lecture room (C1); indoors, wearing different clothes in a different room (C2); outdoors, in a relatively quiet location and wearing the same clothes as in C1 (C3); and outdoors, in a noisy location near a road with traffic and wearing the same clothes as in C2 (C4). Speakers are up to a few meters from the microphone, which is partially covered by the plastic shield of the camera. The pose and appearance of the speakers and the illumination conditions change continuously. High-quality frontal views of the target are rare and often the face of the speaker is not even visible. Figure 2 shows snapshots of these four conditions and Figure 3 shows samples of the detected targets.

Sample audio signals with the related spectrograms are shown in Figure 4. The acoustics change considerably between the indoor conditions, which are quite favourable in C1 and C2, and the outdoor scenario C3. Although quiet, a variety of interfering noise sources are present (e.g. wind blowing into the microphone), which affect the speaker re-identification performance. Finally, C4 is characterised by strong background noise, which at times makes the voice of

[2] The dataset is described in [9] and is available to download here: `http://www.eecs.qmul.ac.uk/~andrea/adaptation.html`

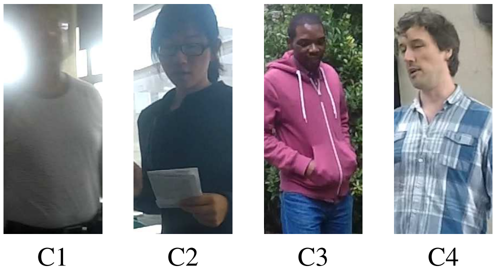


Figure 3: Sample detected targets for the four capturing conditions in the *QM-GoPro* dataset.

the speaker inaudible. Another relevant aspect is the strong low-pass effect, probably due to the plastic shield wrapping the camera, which considerably limits the spectral content even if the audio signals are sampled at 48kHz. Finally, note that in all the cases there is a considerable mismatch (acoustics, noise and language) between the *Q*M-GoPro data and the clean material used for training the I-vector extractor.

Audio streams are at 48kHz, 16 bits. The video resolution is 1920x1080, at 25 frames per second. Each recording is split in 5-second-long segments (skipping the first 8 seconds and the last 5 seconds). A segment, which consists of $N_I = 125$ images and 240000 audio samples, represents a brief interaction between the person wearing the camera and one of the enrolled targets.

A person identity is estimated for each segment (hereafter t is used as segment index). For each condition, the first 3 segments are used for training the two MLP classifiers and the remaining segments are used for testing (approximately 6 to 7 segments per target are available in the test set for each condition). If t_i is the index of the i-th image (or feature vector) of segment t, a target identity is estimated using one of the two modalities or the combined scores:

$$\hat{y}^a(t) = \arg\max_{s \in \mathcal{S}} c^a(s,t) \quad (13)$$

$$\hat{y}^v(t) = \arg\max_{s \in \mathcal{S}} \sum_{i=1}^{N_I} c^v(s,t_i) \quad (14)$$

$$\hat{y}^{av}(t) = \arg\max_{s \in \mathcal{S}} c^{av}(s,t). \quad (15)$$

If K_{corr} is the number of segments whose targets have been correctly recognised and K is the number of test segments, the recognition accuracy is defined as:

$$\text{Acc} = \frac{K_{\text{corr}}}{K} \cdot 100. \quad (16)$$

Depending on how targets appear, the adaptation performance may change considerably (e.g. if a target is present for more than a segment or if "similar" targets are not back-to-back). Therefore, during testing, segments are given to the system in random order, to emulate a realistic scenario where a target is observable only for a short amount of time during the interaction with the person wearing the camera. Results are averaged over 20 random sequences.

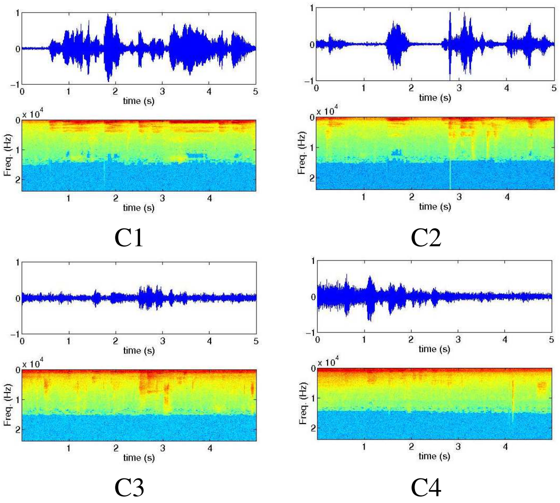
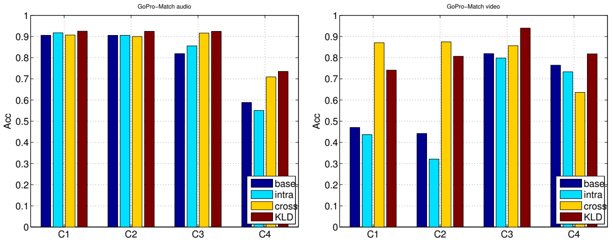


Figure 4: Samples of the audio signals in the four conditions, with related spectrograms.

We evaluate the proposed method when the models are trained in the same environmental conditions considered in the tests (*matched conditions*) and when models are trained in condition C1 (*mismatched conditions*).

5.2. Discussion

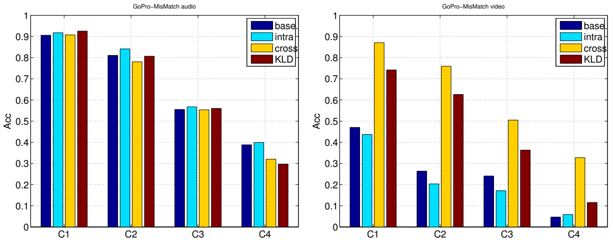
Results in *matched conditions* for the two modalities individually are shown in Figure 5. As expected, the performance in C1 and C2 is good and very similar for the *audio modality* (Figure 5a) and adaptation is therefore unnecessary. The degradation in C3 due to the outdoor environment is improved by “cross” adaptation that brings the performance at the same level as the indoor cases. In C4 the classification accuracy is even lower due to a higher background noise: the proposed model adaptation brings a substantial improvement. In all cases, the “KLD”-based adaptation brings a further improvement to the recognition performance. Note that in this last case the “intra” adaptation deteriorates the performance because the original model is too weak. As for the *video modality*, C1 and C2 are similar challenging scenarios, while in C3 and C4 the target recognition task is simpler (see Figure 5b). The “intra” adaptation always deteriorates the baseline models while the “KLD” adaptation achieves very similar results in all the four conditions. Note that in C4 “cross” adaptation would have failed (audio models are also weak) without the KLD regularisation term, whereas in C1 and C2 the KLD-regularisation limits the improvement provided by the “cross” adaptation, probably due to a too conservative derivation of the KLD



(a) audio

(b) video

Figure 5: Classification accuracy on the QM-GoPro dataset in matched conditions.



(a) audio

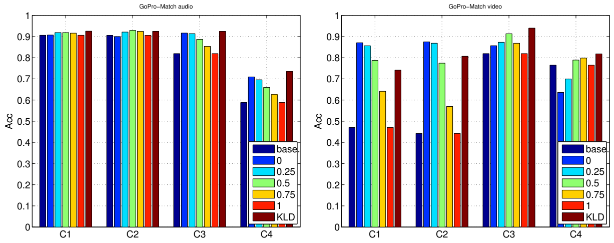
(b) video

Figure 6: Classification accuracy on the QM-GoPro dataset in mismatched conditions.

parameter ρ from the network outputs.

Results in *mismatched conditions* are shown in Figure 6. The baseline performance decreases as the amount of mismatch from C1 increases. In particular, the video performance is very low, therefore adaptation does not help improving the audio models due to unreliable labels. However, excluding the very challenging condition C4, there is no deterioration with respect to the baseline (see Figure 6a). Conversely, the video performance is considerably boosted by the "cross" adaptation. As observed in matched conditions, the KLD term reduces the potential gain, although the resulting performance is above the baseline as well as the "intra" adaptation. This is again related to the way the KLD coefficient is computed, which is probably too conservative. An alternative explanation is that the video models are so weak that adapting is always better, even if labels are not fully reliable. This observation may suggest that ρ should be related not only to the reliability of the labels but also to the degree of mismatch to be compensated for.

Table 1 compares the "cross" adaptation with and without KLD considering the accuracy over 20 randomly generated target sequences. The regularisation term reduces the variance of the classification accuracy, even when the overall average accuracy is lower than what obtained without KLD (see in particular video mismatched conditions).



(a) audio

(b) video

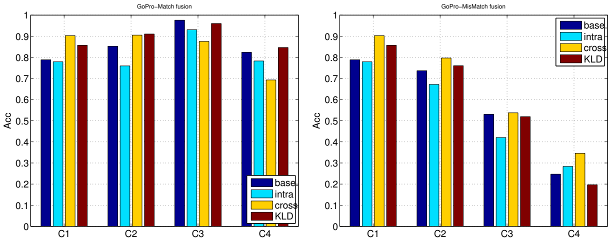
Figure 7: Classification accuracy on the QM-GoPro dataset in matched conditions comparing different values of the KLD regularisation parameter ρ. The same amount of regularisation is used in the whole test set. Performance is compared with the proposed adaptive ρ based on the classifier outputs.

For a better analysis of the behaviour of the KLD regularisation term, Figure 7 reports the classification accuracy in matched conditions using different values of ρ, ranging from 0 (no regularisation) to 1 (no adaptation). Note that in this case ρ is not adapted. The best value for ρ that maximizes the accuracy depends on the conditions and on the accuracy of the other modality. Moreover, adapting ρ on each segment improves the performance with respect to the best static parameters on the audio modality (see Figure 7a). Conversely, the adaptive ρ is not optimum for the video modality in C1 and C2. Instead, the adaptive ρ outperforms the best static parameter in C3 and C4. This is consistent with our claim above that the quality of the audio classifier is probably underestimated in C1 and C2, thus leading to a too conservative adaptation.

We conclude the experimental analysis considering the results of the final system after score fusion (see Figure 8). Note that this fusion strategy is not optimal for the system discussed here as it was developed in [9] to optimise the Equal Error Rate (EER) on a different classifier. Note also that if an effective fusion is used, the overall accuracy is high and the benefits of model adaptation are less evident with respect to the baseline system, especially if the two modalities are always available (as in the majority of the cases in the *QM-GoPro* dataset). As shown in Figure 8a, the performance of the combined system is high in matched conditions. Nevertheless, the "KLD" adaptation is always superior or equivalent to the baseline and the "intra" adaptation. In some cases the KLD term limits the gain achievable, for the same reasons discussed for the video modality alone. However, "cross" adaptation without regularisation does not improve over the baseline in all conditions. Results in mismatched conditions in Figure 8b are in line with what observed for the individual modalities.

Table 1: Average accuracy and standard deviation over the 20 random test sequences for the proposed cross-modal model adaptation with and without KLD regularisation.

Modality	**Adapt**		**Matched**				**Mismatched**			
			C1	C2	C3	C4	C1	C2	C3	C4
Audio	cross	mean	90.71	90.00	91.63	70.94	90.71	78.05	55.36	32.00
		std	2.55	1.91	2.19	3.21	2.55	2.94	4.18	4.86
	KLD	mean	92.53	92.47	92.41	73.47	92.53	80.68	56.02	29.71
		std	1.99	1.53	1.87	3.20	1.99	1.68	1.68	2.49
Video	cross	mean	87.06	87.47	85.66	63.59	87.06	75.95	50.48	32.71
		std	2.75	2.18	2.61	3.13	2.75	2.53	3.26	4.15
	KLD	mean	74.12	80.68	93.98	81.82	74.12	62.53	36.33	11.59
		std	2.38	2.88	1.35	3.01	2.38	1.94	2.00	1.28



(a) matched (b) mismatched

Figure 8: Classification accuracy on the QM-GoPro dataset after score fusion in matched (a) and mismatched (b) conditions.

6. Conclusion

We presented a multi-modal cross-adaptation approach for deep-models and validated it on an audio-visual target re-identification task using data from body worn cameras. Results show the potential of the proposed multi-modal adaptation that, in combination with an appropriate regularisation strategy, leads to a noticeable performance improvement of each single-modal classifier as well as of the late score fusion results.

Future investigations will focus on defining a more robust KLD coefficient ρ and on combining *intra* and *cross* adaptation. Finally, we will investigate the use of more sophisticated mono-modal systems based on a Convolutional Neural Network (CNN) [10] and DNN-based I-vector extraction [26].

References

[1] A. H. Abdelaziz, S. Zeiler, and D. Kolossa. Learning dynamic stream weights for coupled-hmm-based audio-visual speech recognition. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 23(5):863–876, May 2015.

[2] V. Abrash, H. Franco, A. Sankar, and M. Cohen. Connectionist speaker normalization and adaptation. In *Proc. of Eurospeech*, pages 2183–2186, 1995.

[3] P. S. Aleksic and A. K. Katsaggelos. Audio-visual biometrics. *Proc. of the IEEE*, 94(11):2025–2044, 2006.

[4] 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*, pages 1391–1394, 1994.

[5] J. Barker, R. Marxer, E. Vincent, and S. Watanabe. The third CHiME speech separation and recognition challenge: Dataset, task and baselines. In *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 504–511, 2015.

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

[7] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr. Fully-convolutional siamese networks for object tracking. In *Proc. of European Conf. on Computer Vision, Workshops*, pages 850–865, 2016.

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

[9] A. Brutti and A. Cavallaro. Online cross-modal adaptation for audio visual person identification with wearable cameras. *IEEE Trans. on Human-Machine Systems*, 47(1):40–51, Feb 2017.

[10] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016.

[11] C. M. Christoudias, K. Saenko, L.-P. Morency, and T. Darrell. Co-adaptation of audio-visual speech and gesture classifiers. In *Proc. of Int. Conf. on Multimodal Interfaces*, pages 84–91, 2006.

[12] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 886–893 vol. 1, 2005.

[13] 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*, pages 1559–1562, 2009.

[14] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Trans. on Audio, Speech, and Language Processing*, 19(4):788–798, May 2011.

[15] P. Dollár. Piotr's Computer Vision Matlab Toolbox (PMT). http://vision.ucsd.edu/ pdollar/toolbox/doc/index.html.

[16] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 36(8):1532–1545, Aug. 2014.

[17] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(4):743–761, Apr. 2012.

[18] D. Falavigna, M. Matassoni, S. Jalalvand, M. Negri, and M. Turchi. DNN adaptation by automatic quality estimation of ASR hypotheses. *Computer Speech and Language*, 46:585–604, Nov. 2017.

[19] A. Gandhi, A. Sharma, A. Biswas, and O. Deshmukh. Gethrnet: A generalized temporally hybrid recurrent neural network for multimodal information fusion. In *Proc. of European Conf. on Computer Vision*, pages 883–899, 2016.

[20] D. Garcia-Romero, A. McCree, S. Shum, N. Brummer, and C. Vaquero. Unsupervised domain adaptation for I-vector speaker recognition. In *Proc. of Speaker Odissey*, pages 260–264, 2014.

[21] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. D. Mori. Linear hidden transformations for adaptation of hybrid ANN/HMM models. *Speech Communication*, 49(10):827–835, 2007.

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

[23] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, July 2006.

[24] Z. Huang, J. Li, M. Siniscalchi, I. Chen, C. Weng, and C. Lee. Feature space Maximum a Posteriori linear regression for adaptation of Deep Neural Networks. In *Proc. of Interspeech*, pages 2992–2996, 2014.

[25] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(3):226–239, Mar. 1998.

[26] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren. A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, pages 1695–1699, 2014.

[27] 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*, pages 626–633 vol.1, 2003.

[28] B. Li and K. Sim. Comparison of discriminative input and output transformation for speaker adaptation in the hybrid NN/HMM systems. In *Proc. of Interspeech*, pages 526–529, 2010.

[29] S. Lucey, T. Chen, S. Sridharan, and V. Chandran. Integration strategies for audio-visual speech processing: applied to text-dependent speaker recognition. *IEEE Trans. on Multimedia*, 7(3):495–506, June 2005.

[30] A. Ma, J. Li, P. Yuen, and P. Li. Cross-domain person reidentification using domain adaptation ranking SVMs. *IEEE Trans. on Image Processing*, 24(5):1599–1613, May 2015.

[31] Y. Miao and F. Metze. Improving low-resource CD-DNN-HMM using dropout and multilingual DNN training. In *Proc. of Interspeech*, pages 2237–2241, 2013.

[32] J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals, and T. Robinson. Speaker adaptation for hybrid HMM-ANN continuous speech recognition system. In *Proc. of Eurospeech*, pages 2171–2174, 1995.

[33] A. Preti and J.-F. Bonastre. Unsupervised model adaptation for speaker verification. In *Proc. of Interspeech*, pages 1–4, 2006.

[34] F. Richardson, D. Reynolds, and N. Dehak. Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters*, 22(10):1671–1675, Oct 2015.

[35] F. Seide, G. Li, X. Chen, and D. Yu. Feature engineering in context-dependent deep neural networks for conversational speech transcription. In *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 24–29, 2011.

[36] S. Shivappa, M. Trivedi, and B. Rao. Audiovisual information fusion in human computer interfaces and intelligent environments: A survey. *Proc. of the IEEE*, 98(10), Oct. 2010.

[37] S. Siniscalchi, J. Li, and C. Lee. Hermitian polynomial for speaker adaptation of connectionist speech recognition systems. *IEEE Trans. on Audio, Speech, and Language Processing*, 21(10):2152–2161, 2013.

[38] R. Stiefelhagen, K. Bernardin, R. Bowers, R. Rose, M. Michel, and J. Garofolo. The CLEAR 2007 Evaluation Multimodal Technologies for Perception of Humans. In Stiefelhagen, Bowers, and Fiscus, editors, *Multimodal Technologies for Perception of Humans*, LNCS, chapter 1, pages 3–34. Springer, Berlin, Heidelberg, 2008.

[39] S. Tamura, H. Ninomiya, N. Kitaoka, S. Osuga, Y. Iribe, K. Takeda, and S. Hayamizu. Investigation of DNN-based audio-visual speech recognition. *IEICE Trans. on Information and Systems*, 99:2444–2451, 2016.

[40] R. Vezzani, D. Baltieri, and R. Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys*, 46(2):29:1–29:37, Dec. 2013.

[41] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong. Adaptation of context-dependent Deep Neural Networks for automatic speech recognition. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, pages 366–369, 2012.

[42] D. Yu, K. Yao, H. Su, G. Li, and F. Seide. KL-Divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, pages 7893–7897, 2013.

[43] 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*, pages 1–8, 2015.