ON THE UTILITY OF CANONICAL CORRELATION ANALYSIS FOR DOMAIN ADAPTATION IN MULTI-VIEW HEADPOSE ESTIMATION

Anoop K.R.[1], *Ramanathan Subramanian*[2], *Vassilios Vonikakis*[2], *K.R. Ramakrishnan*[1] *and Stefan Winkler*[2]

[1] Dept. of Electrical Engineering, Indian Institute of Science, Bengaluru
[2] Advanced Digital Sciences Center (ADSC), University of Illinois at Urbana-Champaign, Singapore

ABSTRACT

The utility of **canonical correlation analysis** (CCA) for domain adaptation (DA) in the context of multi-view head pose estimation is examined in this work. We consider the three problems studied in [1], where different DA approaches are explored to transfer head pose-related knowledge from an extensively labeled *source* dataset to a sparsely labeled *target* set, whose attributes are vastly different from the *source*. CCA is found to benefit DA for all the three problems, and the use of a covariance profile-based **diagonality score** (DS) also improves classification performance with respect to a nearest neighbor (NN) classifier.

Index Terms— Canonical Correlation Analysis, Domain Adaptation, Head pose classification, Diagonality score

1. INTRODUCTION

Due to the extensive difficulty and cost of annotating large datasets, **domain adaptation** (DA) or transfer learning techniques, which allow for adapting models trained on existing datasets to novel ones, have become very popular recently. Traditional learning algorithms require feature-wise consistent *target* (or test) data for achieving good performance– this requirement is violated when the *target* data is sufficiently different in nature with respect to the *source* (or training) data. In such cases, transfer learning adapts knowledge acquired from the *source* by incorporating *target*-specific information learned from a few labeled *target* examples.

We specifically examine DA in the context of multi-view coarse head pose estimation (or head pose classification) in this paper. Head pose estimation from surveillance video is particularly useful for security and behavior analysis, but has been shown to be challenging even with multi-view systems in [2]. Three DA approaches for determining the head pose class of a person (target) captured by four surveillance cameras are evaluated in [1]. Upon training models with CLEAR (*source*) images where targets rotate in-place and exhibit a roughly frontal head-tilt as in social interactions, the authors attempt head pose classification on the DPOSE (*target*) dataset specifically considering three problems as illustrated in Fig.1. P1 denotes the case where DPOSE targets also rotate in-place while exhibiting a larger range of head tilts than in the CLEAR images. P2 represents the situation where DPOSE targets are moving, but exhibiting head-tilts similar to CLEAR. P3 combines P1 and P2, and involves moving targets exhibiting a large range of head-tilts (as in a museum or supermarket). In addition to the fundamental difficulty in determining head pose from low-resolution face images, traditional learning algorithms trained on CLEAR perform poorly on DPOSE due to differences in camera geometry and lighting, and mainly because the facial appearance of DPOSE targets varies under motion due to changing camera perspective and scale. In contrast, CLEAR targets remain stationary ensuring greater consistency in appearance.

Three DA approaches are proposed in [1] to solve P1-P3. For P1, an adaptation of the inductive transfer learning algorithm [3] is proposed so that misclassified *target* samples are preferentially learned in subsequent iterations. Transferable distance functions are proposed for P2 and P3, where patch weights signifying importance of face patches for pose estimation are learned from *source* data, and adapted to the *target* by accounting for appearance distortions arising from perspective and scale changes. A nearest neighbor (NN) classifier is then employed to classify *target* samples.

Given that DA is necessitated by the *source* and *target* data having different feature distributions, projecting both onto a correlated sub-space to make them feature-wise consistent can enhance transfer learning. Canonical correlation analysis (CCA) is useful to this end [4], as it enables discovery of linear relationships between two feature sets. We show that CCA improves DA performance for P1-P3 via extensive experiments. Also, in the context of transferable distance learning, the NN classifier chooses the closest neighbor among a (typically small) set of class-specific prototypes. Deriving *class-representative signatures* instead would be beneficial. Covariance profiles (CP) [5] enable efficient description of an object *family*, and the closeness of a novel object to this family can be computed using the diagonality score (DS). We also show that employing the diagonality

This study is supported by the research grant for the Human-Centered Cyber-physical Systems Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR).

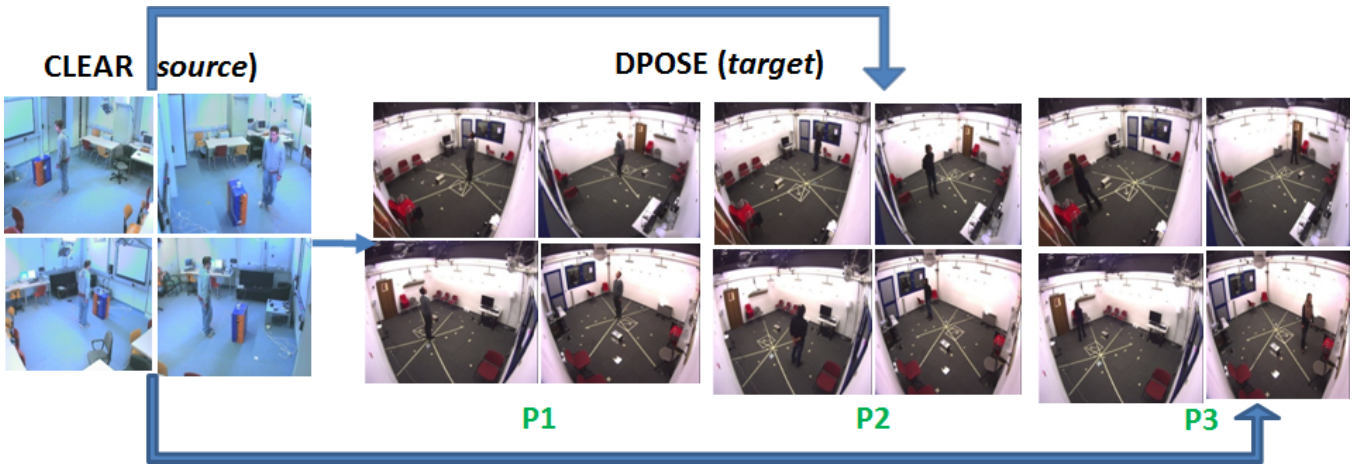


Fig. 1. Illustration of the three problems (P1, P2, P3) studied in [1]. 4-view images from CLEAR and DPOSE are shown two-by-two. Figure is best viewed in color and under zoom.

score is advantageous and enhances DA-based head pose classification performance with respect to the NN classifier.

2. RELATED WORK

We now briefly review related work on (i) head pose estimation (HPE) from surveillance videos and related domain adaptation approaches, and (ii) the use of CCA and covariance profiles for computer vision applications.

HPE from low resolution videos: A Kullback Leibler distance-based face descriptor is proposed in [6], but is outperformed by the array-of-covariances (ARCO) [7], which is also robust to occlusions, scale and illumination changes. Recent works [8,9] have explored the use of weak labels (*e.g.* motion direction for head pose) for robust HPE with unlabeled data. Nevertheless, all these works attempt monocular HPE. Multi-view HPE has only been addressed by a handful of algorithms. A particle filter is combined with two neural networks for estimating head pan and tilt in [10]. Multi-view HPE under target motion [11] is achieved by mapping multiple face images onto a textured 3D model, and determining the face location in the unfolded texture map. An unsupervised approach to tackle face appearance variations under motion is proposed in [12], where spectral clustering is employed to segment the scene into regions, and region-specific head pose classifiers are learned. Multi-task learning for estimating head pose under motion is proposed in [13].

As obtaining labeled head pose data in surveillance settings is difficult and expensive, few works have explored the use of DA techniques for HPE on novel datasets. An adaptive multiple kernel learning-based active DA framework is proposed in [14]. An extensive discussion on how DA can achieve efficient multi-view head pose classification in novel scenarios is presented in [1].

CCA and CP in vision applications: The correlated subspace derived from CCA is exploited for transfer learning, and is applied to a number of vision problems including action recognition in [15]. A methodology to employ CCA for domain-adaptive head pose classification is presented in [16]. Some works have also focused on representing and comparing a family of objects– principal angles [17] are popular in this respect. Covariance profiles are shown to be effective for object track clustering and face recognition in [5].

3. PROPOSED FRAMEWORK

Fig.2 presents an overview of the proposed framework. In [1], ARCO [7] is modified via Tradaboost [3] to develop the ARCO-Xboost classifier. For P2, upon dividing the 4-view face image into overlapping 8×8 patches, weights denoting saliency of these patches for HPE are learned from the *source* data. These weights are then modulated to the *target* dataset incorporating patch reliability scores encoding facial appearance distortion due to motion from a *reference* scene location. Finally, the *target* (test) sample label is assigned on determining its nearest neighbor based on the weighted distance between corresponding patches. A similar approach is adopted for P3, where patch weights are decomposed into hyperfeatures which encode the patch saliency and parameters P, which are directly transferable from the *source* to the *target* and are learned from *source* data. These two weighted distance approaches are denoted as WD (see [1]) in Fig.2.

We consider the same scenario as in [1], where the *source* data comprises only images corresponding to a frontal head-tilt, while the *target* dataset comprises a much larger range of head tilts. More specifically, the *source* data is discretized into 8 classes each denoting a head pan range of 45°, while the target data is discretized into 24 classes (8 pan $\times$ 3 namely, upward, frontal and downward tilt ranges). For P1 and P3, our objective is to assign a class label $C \in [1..24]$ for each test sample, while there are only eight possible classes for P2 as only those *target* instances consistent with the *source*

in terms of head pose are considered here. In the proposed framework, we perform CCA on the *source* and *target* features to derive a correlated subspace, and feed the *source* and *target* features **projected** onto the correlated subspace to the DA algorithms for P1-P3. For P2 and P3, we also examine if the use of the diagonality score with covariance features can improve performance with respect to the NN classifier. Brief descriptions of CCA and CP are as follows.

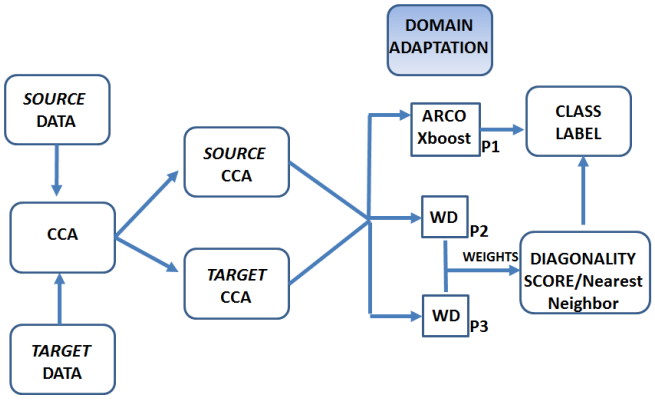


Fig. 2. Overview of the proposed framework.

3.1. Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) [18] is a technique used to analyze the linear relationship between two multi-dimensional variables. CCA seeks to find two basis vectors such that the correlation between the two variables is mutually maximized. CCA can also be viewed as a measure of similarity between two variables, and the two variables involved can be of different dimensions. A probabilistic interpretation of CCA as a latent model is given in [19]. Most importantly, CCA places the fewest restrictions on the types of data on which it operates.

Let X and Y be two random variables. Consider a set of n samples $x_i \in \mathcal{X} \subseteq \Re^{n1}$ and $y_i \in \mathcal{Y} \subseteq \Re^{n2}$ forming n training samples $\{(x_i, y_i)\}_{i=1}^{n}$. CCA seeks to find vectors a and b such that the random variables $a'X$ and $b'Y$ are correlated maximally. Let $\mathbf{X} = [x_1, ..., x_n] \in \Re^{n1 \times n}$ and $\mathbf{Y} = [y_1, ..., y_n] \in \Re^{n2 \times n}$. We assume that the data is centered i.e. $E[X] = E[Y] = 0$. Let $\Sigma_{XX} = E[\mathbf{XX}']$, $\Sigma_{YY} = E[\mathbf{YY}']$ and $\Sigma_{XY} = E[\mathbf{XY}']$. The optimization problem is thus as given below

$$\underset{a \in \Re^{n1}, b \in \Re^{n2}}{\operatorname{argmax}} \frac{a' \Sigma_{XY} b}{\sqrt{(a \Sigma_{XX} a)(b \Sigma_{YY} b)}} \tag{1}$$

Note that the above problem is invariant to scaling of a and b. Thus, the maximization problem can be redefined with the following constraints $a' \Sigma_{XX} a = 1$ and $b' \Sigma_{YY} b = 1$. The optimization problem in Eq. (1) can be written as a constrained optimization problem as below

$$\underset{a \in \Re^{n1}, b \in \Re^{n2}}{\operatorname{argmax}} \frac{a' \Sigma_{XY} a}{\sqrt{(a' \Sigma_{XX} a)(b' \Sigma_{YY} b)}} \tag{2}$$

$$\text{s.t. } a' \Sigma_{XX} a = 1 \text{ and } b' \Sigma_{YY} b = 1 \tag{3}$$

The canonical correlations can be found by solving the generalized eigenvalue problem below to obtain d pairs of eigenvectors for each random variable to be projected:

$$\Sigma_{XX} - 1 \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX} a = \lambda^2 a$$

$$\text{and } b = \frac{\Sigma_{YY}^{-1} \Sigma_{YX} a}{\lambda}$$

3.2. Covariance Profiles

Covariance profile (CP) is a descriptor for an object set described by a covariance matrix. CP descriptors are used for video object track clustering and face recognition in [5]. Given a set of similar objects $T = \{C_1, C_2, \ldots, C_N\}$ where each C_i is a covariance descriptor [20], CP attempts to find a matrix which captures some notion of similarity among these covariance descriptors. Mathematically, the problem is to obtain vectors $\beta_1, \beta_2, \ldots, \beta_d$, where d is the number of columns of the C_i's, such that

$$C_i = \sum_j \lambda_{ij} \beta_j \beta_j' \tag{4}$$

The matrix P, with vectors $\{\beta_j\}$ as its columns, *jointly diagonalizes* the individual C_i matrices, *i.e.* $\Lambda_i = P' C_i P$ is a diagonal matrix with entries λ_{ij}. This matrix is defined as the **covariance profile** for the set.

To estimate a CP of a given set, we perform *approximate joint diagonalization* using Pham's algorithm [21] designed for joint diagonalization of positive definite Hermitian matrices. Also, CP can be used as a similarity measure– the closeness of an object with covariance descriptor C to a family T is computed using the *diagonality score* (DS). This value decreases with decreasing Frobenius norm of the off-diagonal elements in $P'CP$ and is 0 if and only if it is fully diagonal [22]. Thus, given a P for a family T, the diagonality measure of C is given by Eq. (5). A sufficiently small diagonality score indicates that C is almost perfectly diagonalized by P, and hence likely to belong to the family T.

$$\log\left(\frac{\det(\operatorname{diag}(P'CP))}{\det(P'CP)}\right) \tag{5}$$

4. EXPERIMENTS

In this section, we examine the utility of CCA and CP for the three DA problems P1–P3 described in Sec.3. We use CLEAR [23] as the *source* dataset, and DPOSE [24] as the *target*. As mentioned earlier, *source* dataset comprises only images with frontal head-tilt (8 classes), while the *target* dataset also includes images with upward and downward

head-tilts (24 classes in total). Classification accuracies on the *target* are presented for *source* (classes C1-C8, denoted acc_{src}), *non-source* (C9-C24, denoted acc_{nsrc}) and *all* (acc_{all}) classes for P1 and P3. For all experiments, 300 samples/class were used in the *source* training set, while 5 samples/class were used in the *target* training set. Covariance (cov), LBP and HoG features were computed as in [1].

Table 1 presents the impact of CCA on the DA framework for P1. Classification results on combining CCA with the ARCO and ARCO-Xboost classifiers are compared with the corresponding baselines. Projecting both the *source* and *target* data onto the correlated sub-space considerably improves classification performance for *source* classes with respect to ARCO, while there is a marginal improvement over ARCO-Xboost. This in turn improves overall classification accuracy, even if the performance of CCA+ARCO/CCA+ARCO-Xboost does not improve with respect to ARCO/ARCO-Xboost for non-source classes as very few training examples from these classes are used for model synthesis.

P2 results are shown in Table 2. Here, WD+CCA performs better than WD considering the different scene regions (as in [1]), while mean accuracy increases by 5%. Improvement on combining diagonalization score (DS) with WD is marginal. However, CP is a concise representation of a training class. Fig. 3(a) shows variation in classification accuracy with different CCA-subspace dimensions. We observe that good accuracies are observed even with very low dimensionality. Also, WD+CCA (with nearest *target* neighbor denoted as NN_{tgt}) outperforms CCA (NN_{tgt}), which evidences the need for DA. Also, WD+CCA (NN_{src} denoting nearest *source* neighbor) performs almost as well as WD+CCA (NN_{tgt}). WD+CCA (NN_{tgt}) decreases with increasing dimensionality, while WD+CCA (NN_{src}) remains constant which could be due to the large number of *source* samples available for comparison.

The results for P3 is shown in Table 3. In [16] better performance of WD+CCA over WD is shown. Also, WD+DS performs better than WD+CCA by a minimum of about 3%. WD+DS performs better than WD+CCA for *source* classes, which contributes to better classification. Fig. 3(b) presents variation in overall classification accuracy with varying training size. DS_{tgt} (only DS on *target* prototypes without DA) performs better than NN_{tgt} while WD+DS performs best with increasing dimensionality due to compact representation of the training samples via CP.

5. CONCLUSION

This work examines the utility of canonical correlation analysis and covariance profile for domain adaptive multi-view headpose classification. Empirical results confirm that CCA and the use of the diagonalization score (DS) can improve classification performance for DA problems P1-P3. Integrating CCA and CP, which is computationally demanding, to develop a unified DA framework is left to future work.

Table 1. **P1:** Performance comparison for static targets. Mean *target* classification accuracy reported over 24 classes.

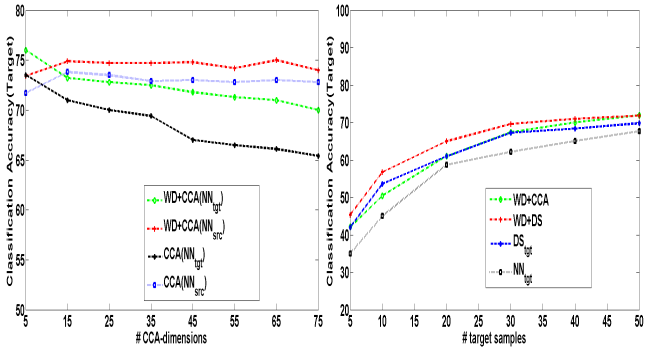
	ARCO [7]	CCA+ARCO	ARCO-Xboost [1]	CCA+ARCO-Xboost
acc_{src}	41	56.1	73.8	74.5
acc_{nsrc}	7	2	31.1	31
acc_{all}	18	18.7	45.4	**46**

Table 2. **P2:** Performance comparison for 8-class head-pan classification under target motion. The room is divided into 4 quadrants (R1-R4). NN classifier with Targets is considered.

	WD $Cov\ (d=12)$	WD LBP	WD+CCA $Cov\ (d=12)$	WD+DS $Cov\ (d=12)$
R1	69.8	**74.7**	74.2	74.1
R2	72.4	77.6	**81.3**	78.7
R3	63	66.9	**73**	66.1
R4	62.5	64.5	**74.9**	65.9
Regions Average	66.9	70.9	**75.9**	71.2

Table 3. **P3:** Classification accuracies for *source* (C1–C8), *non-source* (C9–C24) and *all* classes with freely moving targets. DS results available only with covariance features.

$\boldsymbol{acc}_{src}$	$Cov+LBP$	$Cov+HoG$	$LBP+HoG$
WD+CCA [16]	36.7 ± 0.7	35.1 ± 0.6	34.6 ± 0.7
WD + DS	41.2 ± 0.9	$\mathbf{42.8} \pm 1$	-
DS_{tgt}	37.5 ± 0.7	37.5 ± 0.7	-
NN_{tgt}	31.2 ± 0.8	31.2 ± 0.8	32.4 ± 1
acc_{nsrc}	$Cov+LBP$	$Cov+HoG$	$LBP+HoG$
WD+CCA [16]	48.8 ± 0.8	47.1 ± 0.5	42.3 ± 0.9
WD + DS	49.5 ± 0.9	$\mathbf{49.7} \pm 1$	-
DS_{tgt}	46.7 ± 0.6	46.7 ± 0.6	-
NN_{tgt}	38.8 ± 0.8	38.8 ± 0.8	40.2 ± 0.8
acc_{all}	$Cov+LBP$	$Cov+HoG$	$LBP+HoG$
WD+CCA [16]	42.1 ± 0.8	40.5 ± 0.6	38.9 ± 0.8
WD + DS	45.3 ± 1	$\mathbf{46.2} \pm 0.8$	-
DS_{tgt}	42 ± 1	42 ± 1	-
NN_{tgt}	35.8 ± 0.9	35.8 ± 0.9	36.3 ± 0.9



(a) (b)

Fig. 3. (a) Variation in overall classification accuracy with increasing CCA dimension for P2. (b) Variation in overall classification accuracy with increase in *target* training examples for problem P3.

6. REFERENCES

[1] Anoop Kolar Rajagopal, Ramanathan Subramanian, Elisa Ricci, Radu L Vieriu, Oswald Lanz, Nicu Sebe, and Kalpathi R. Ramakrishnan, "Exploring transfer learning approaches for head pose classification from multi-view surveillance images," *IJCV*, vol. 109, no. 1-2, pp. 146–167, 2014.

[2] Ramanathan Subramanian, Yan Yan, Jacopo Staiano, Oswald Lanz, and Nicu Sebe, "On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions," in *ICMI*, 2013.

[3] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu, "Boosting for transfer learning," in *ICML*, 2007, pp. 193–200.

[4] Yi-Ren Yeh, Chun-Hao Huang, and Yu-Chiang Frank Wang, "Heterogeneous domain adaptation and classification by exploiting the correlation subspace," *IEEE Trans. on Image Processing*, vol. 23, no. 5, pp. 2009–2018, 2014.

[5] Anoop KR, Adway Mitra, Ujwal Bonde, Chiranjib Bhattacharyya, and KR Ramakrishnan, "Covariance profiles: A signature representation for object sets," in *ICPR*, 2012, pp. 2541–2544.

[6] Javier Orozco, Shaogang Gong, and Tao Xiang, "Head pose classification in crowded scenes," in *BMVC*, 2009.

[7] Diego Tosato, Michela Farenzena, Marco Cristani, Mauro Spera, and Vittorio Murino, "Multi-class classification on Riemannian manifolds for video surveillance," in *ECCV*, 2010.

[8] Cheng Chen and Jean-Marc Odobez, "We are not contortionists: coupled adaptive learning for head and body orientation estimation in surveillance video," in *CVPR*, 2012.

[9] B. Benfold and I. Reid, "Unsupervised learning of a scene-specific coarse gaze estimator," in *ICCV*, 2011.

[10] Michael Voit and Rainer Stiefelhagen, "A system for probabilistic joint 3D head tracking and pose estimation in low-resolution, multi-view environments," in *Computer Vision Systems*, 2009.

[11] Xenophon Zabulis, Thomas Sarmis, and Antonis A. Argyros, "3D head pose estimation from multiple distant views," in *BMVC*, 2009.

[12] Isarun Chamveha, Yusuke Sugano, Daisuke Sugimura, Teera Siriteerakul, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto, "Head direction estimation from low resolution images with scene adaptation.," *CVIU*, vol. 117, no. 10, pp. 1502–1511, 2013.

[13] Yan Yan, Elisa Ricci, Ramanathan Subramanian, Oswald Lanz, and Nicu Sebe, "No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion," in *ICCV*, 2013.

[14] Yan Yan, R. Subramanian, O. Lanz, and N. Sebe, "Active transfer learning for multi-view head-pose classification," in *ICPR*, 2012.

[15] Yi-Ren Yeh, Chun-Hao Huang, and Y.-C.F. Wang, "Heterogeneous domain adaptation and classification by exploiting the correlation subspace," *Image Processing, IEEE Trans. on*, vol. 23, no. 5, pp. 2009–2018, 2014.

[16] Anoop K.R. and K.R.Ramakrishnan, "Domain adaptation on correlated subspace for unseen multiview head-pose classification," in *ICVGIP*, 2014.

[17] O. Yamaguchi, Fukui, and Maeda, "Face recognition using temporal image sequence," in *FGR*, 1998, pp. 318–323.

[18] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 312–377, 1936.

[19] Francis R Bach and Michael I Jordan, "A probabilistic interpretation of canonical correlation analysis," *Technical Report 688, Department of Statistics, University of California, Berkeley*, 2005.

[20] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," *ECCV*, pp. 589–600, 2006.

[21] D. T. Pham, "Joint Approximate Diagonalization of Positive Definite Hermitian matrices," *SIAM J. Matrix Anal. Appl.*, vol. 22, no. 4, 2001.

[22] Bernhard N. Flury, "Common principal components and related multivariate models," *John Wiley and Sons*, 1988.

[23] Rainer Stiefelhagen, Rachel Bowers, and Jonathan Fiscus, *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*, vol. 4625, Springer, 2008.

[24] A. Rajagopal, Ramanathan Subramanian, Radu L. Vieriu, Elisa Ricci, Oswald Lanz, Nicu Sebe, and Kalpathi Ramakrishnan, "An adaptation framework for head pose estimation in dynamic multi-view scenarios," *ACCV*, pp. 652–666, 2012.