COUPLED DICTIONARIES FOR THERMAL TO VISIBLE FACE RECOGNITION

Christopher Reale[⋆†] *Nasser M. Nasrabadi*[†] *Rama Chellappa*[⋆]

[⋆]Department of Electrical and Computer Engineering and the
Center for Automation Research, UMIACS, University of Maryland, College Park, MD
[†] U.S. Army Research Laboratory, Adelphi, MD

ABSTRACT

Thermal to visible face recognition is the problem of identifying a thermal infrared (IR) face image given a gallery of visible light face images. We attempt to solve this problem by learning coupled dictionaries to represent the two domains. The dictionaries provide a sparse representation which transforms the data into a single, domain-independent, latent space. We formulate the dictionary learning problem as a bi-level optimization problem and perform a stochastic gradient descent on the dictionaries to solve it. We present experimental results demonstrating the effectiveness of our approach.

1. INTRODUCTION

The ability to identify a person by a thermal image of their face is primarily of interest to law enforcement agencies. Authorities may want to identify a suspect from images taken at night when there may not be enough ambient light to support visible light imagery. Visible light imagery requires ambient light because it is mainly reflective with respect to humans. Thermal IR is mainly emissive and thus does not depend on a certain amount of ambient light being present.

Consider the general face recognition problem. The objective is to determine the identity of a probe face image given a database of gallery face images. Over the past two decades there have been an abundance of algorithms proposed to solve this problem. Most attempts use some sort of machine learning algorithm to train a classifier on the gallery and then apply the classifier to the probe images to identify them. While this is generally a good approach to take, virtually all machine learning algorithms operate under the assumption that the training and testing data are drawn from the same underlying modality. This may hold for general face recognition, but by definition it is not the case for thermal to visible face recognition. Classifying thermal facial images using a classifier trained on visible light images will result in poor performance.

One way to bridge the discrepancy between the two domains is to transform images from both domains into a single common space. Choi et al. [1] attempt to use image filtering and feature extraction to achieve this goal for thermal to visible face recognition, but it is not clear that their method is optimal. In many cases (including our's) a specific transformation is not immediately apparent. In this paper we take a data-driven approach to learn a coupled transformation that maps thermal and visible images to a common domain. We attempt to leverage a training set containing pairs of face images (each pair contains a single subject viewed in both domains) to perform this task. It is important to note that the set of subjects in the training set and the set of subjects in the probe images are *disjoint* and so the training set can only be used to learn the relationship between the domains. We adopt this convention to better mimic real-world scenarios.

While the face images of a subject are obtained by different sensing modalities, they both portray data originating from the same underlying physical object. We attempt to extract latent information that is common to the two domains by using coupled dictionaries [2]. We train the domain dictionaries such that their atoms capture phenomena that are observable across domains. The presence and markedness of these phenomena in a sample are encoded by the sample's sparse coefficients. These coefficients serve as an underlying, domain-independent representation of a subject. We refer to the space in which the coefficients lie as a domain-insensitive latent subspace.

The paper is organized as follows. In section 2 we give a brief overview of recent work that is relevant to this paper. In section 3 we give some background on sparse coding and describe the formulation we use for the problem. In section 4 we describe the optimization algorithm used as well as other implementation details. In section 5 we describe experiments we performed to test our method and give their results.

2. RELATED WORK

Sparse coding is a very heavily researched topic, particularly in the computer vision community. Wright et al. [3] were the first to use sparse coding in a computer vision setting for face recognition. Since then there has been an explosion of work related to this subject. Many have proposed algorithms to learn a dictionary rather than using training samples as atoms [4]. While in general, most dictionaries are used for reconstructive purposes, they can also be trained to support alternative tasks.

Coupled dictionaries have also been used recently for computer vision and image processing applications. Yang et al. [5] first introduced coupled dictionaries by concatenating the features from two domains and training a dictionary using a standard single-domain algorithm. They improved upon their method by applying bilevel constraints for one of the domains in [6] and both domains in [2]. Wang et al. [7] allowed the sparse codes for the two domains to be approximately related by a linear mapping rather than taking the same values. Jia et al. [8] assumed that the sample's sparse representations in the two domains must have a common support but do not necessarily have to be equal. He et al. [9] utilized a Beta process prior to train coupled dictionaries. While most of this work focused on reconstruction tasks (super-resolution, face synthesis), Zheng et al. [10] used coupled dictionaries for cross-view action recognition.

While coupled dictionary learning has received significant attention, thermal to visible face recognition has not been as heavily researched. Bourlai et al. [11] experimented with different methods for image preprocessing and feature extraction. Choi et al. [1] evaluated the use of Partial Least Squares in the context of thermal to visible face recognition. Klare and Jain [12] use kernel similarities to a fixed training set as features for matching. Ghiass et al. provide a more extensive literature review [13].

3. PROPOSED METHOD

3.1. Sparse Coding

The goal of sparse coding is to compactly represent a signal $\mathbf{x} \in \mathbb{R}^d$ as a linear combination of a small number of basis atoms from an over-complete dictionary $\mathbf{D} \in \mathbb{R}^{d\times K}(d < K)$. An important aspect of sparse coding is having a dictionary with respect to which the targeted signals are sparse. While some analytic bases (e.g. Fourier basis, Wavelet basis) can be used for certain types of signals, dictionaries learned from a training set of signals tend to perform better. One popular method of learning a dictionary is by solving the following optimization problem. Given a training set $\{\mathbf{x}_i\}_{i=1}^N$ and regularizing parameter λ

$$\begin{aligned} &\min_{\mathbf{D},\{\boldsymbol{\alpha}_i\}_{i=1}^N} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_i\|_1 \\ &\text{s.t.} \|\mathbf{D}(:,k)\|_2 \leq 1, \forall k \in \{1, 2, \ldots, K\}, \end{aligned} \tag{1}$$

where the ℓ_1 norm serves to promote sparsity in $\boldsymbol{\alpha}_i \in \mathbb{R}^K$, the basis coefficients of sample i. This problem is convex in $\mathbf{D}$ and convex in $\boldsymbol{\alpha}$, but not in both. Approaches for solving this formulation typically find a local minimum by alternately minimizing over $\mathbf{D}$ and $\boldsymbol{\alpha}$.

3.2. Bilevel Coupled Sparse Coding

Coupled sparse coding attempts to accomplish the same goal as sparse coding, but across multiple domains. That is, it attempts to create a compact information preserving representation for signals that span or can be viewed in multiple domains. As with sparse coding in a single domain, overcomplete dictionaries learned from data tend to outperform analytic bases. While there are many ways to learn coupled dictionaries, here we follow the approach in [2] and solve the following optimization problem to learn coupled dictionaries $\mathbf{D}_x$ and $\mathbf{D}_y$ for signals $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$

$$\begin{aligned} &\min_{\mathbf{D}_x,\mathbf{D}_y} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{D}_x\mathbf{z}_i^x\|_2^2 + \|\mathbf{y}_i - \mathbf{D}_y\mathbf{z}_i^y\|_2^2 + \gamma\|\mathbf{z}_i^x - \mathbf{z}_i^y\|_2^2 \\ &\text{s.t. } \mathbf{z}_i^x = \arg\min_{\boldsymbol{\alpha}} \|\mathbf{x}_i - \mathbf{D}_x\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1, \forall i \\ &\quad \mathbf{z}_i^y = \arg\min_{\boldsymbol{\alpha}} \|\mathbf{y}_i - \mathbf{D}_y\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1, \forall i \\ &\quad \|\mathbf{D}_x(:,k)\|_2 \leq 1, \forall k \in \{1, 2, \ldots, K\} \\ &\quad \|\mathbf{D}_y(:,k)\|_2 \leq 1, \forall k \in \{1, 2, \ldots, K\}. \end{aligned} \tag{2}$$

Due to the bilevel nature of this optimization problem (i.e. some of the constraints contain optimization problems themselves) we term the generated dictionaries as bilevel coupled dictionaries. As with single domain dictionaries, coupled dictionaries are trained to minimize the reconstruction error of the original signal(s). This is done through the first two terms of the objective function: $\|\mathbf{x}_i - \mathbf{D}_x\mathbf{z}_i^x\|_2^2$ and $\|\mathbf{y}_i - \mathbf{D}_y\mathbf{z}_i^y\|_2^2$. Note that in this formulation, the sparse coefficients are explicitly calculated from a single domain. This is important for our application because as previously discussed, we will not have data available from both domains for probe and gallery samples at test time. The trade-off for this is that the sparse coefficients for the two domains will not be exactly the same. We can minimize this problem through the third term in the objective function: $\|\mathbf{z}_i^x - \mathbf{z}_i^y\|_2^2$. The trade-off between the reconstructive ability and sparse coefficient similarity can be tweaked by altering the balance parameter γ.

4. IMPLEMENTATION

4.1. Patch Based Approach

Due to the high dimensionality of our data and the relatively few number of training samples, it is not possible to train a generalized dictionary pair for the entire feature space. Instead, we train dictionary pairs on features extracted from patches of the face images rather than the whole images. Not only does this reduce the dimension of the feature space, it also increases the number of training samples as each face image provides many patches.

While the dictionaries we learn encode patches, we still want to train a classifier that operates on an entire face image. We form a grid of overlapping patches across the face and concatenate their sparse coefficients for classification purposes.

4.2. Algorithm

Like many dictionary learning formulations, our optimization problem (2) is highly non-convex. To get around this, we first use K-SVD [4] to initialize the dictionaries. We initialize them jointly by concatenating the features of the two domains to solve the following optimization problem as in [5]

$$\begin{aligned} \min_{\mathbf{D}_x,\mathbf{D}_y,\boldsymbol{\alpha}} \sum_{i=1}^{N} \|\mathbf{x}_i - \mathbf{D}_x\boldsymbol{\alpha}_i\|_2^2 + \|\mathbf{y}_i - \mathbf{D}_y\boldsymbol{\alpha}_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_i\|_1 \\ \text{s.t.}\|\mathbf{D}_{\mathbf{x}}(:,k)\|_2 \leq 1, \forall k \in \{1,2,\ldots,K\} \qquad (3) \\ \|\mathbf{D}_{\mathbf{y}}(:,k)\|_2 \leq 1, \forall k \in \{1,2,\ldots,K\}. \end{aligned}$$

We then perform a stochastic gradient descent on the dictionaries to find a local minimum of (2). Each iteration of the descent has two main steps. We first estimate the gradient of the objective function with respect to the dictionaries on a subset of the training samples. In the second step we update the dictionaries based on the gradient and normalize the atoms of the dictionaries to have unit norm (See Algorithm 1).

Algorithm 1 Gradient Descent

```
1:  $t \Leftarrow 1$
2:  repeat
3:      Randomly choose $S \subset \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$
4:      for $\{\mathbf{x}_i, \mathbf{y}_i\} \in S$ do
5:          $(\nabla L_i)_{\mathbf{D}_x} \Leftarrow \frac{\partial L_i}{\partial \mathbf{D}_x} + \frac{\partial L_i}{\partial \mathbf{z}_i^x} \cdot \frac{\partial \mathbf{z}_i^x}{\partial \mathbf{D}_x}$
6:          $(\nabla L_i)_{\mathbf{D}_y} \Leftarrow \frac{\partial L_i}{\partial \mathbf{D}_y} + \frac{\partial L_i}{\partial \mathbf{z}_i^y} \cdot \frac{\partial \mathbf{z}_i^y}{\partial \mathbf{D}_y}$
7:      end for
8:      $(\nabla L)_{\mathbf{D}_x} \Leftarrow \frac{1}{|S|} \sum_i (\nabla L_i)_{\mathbf{D}_x}$
9:      $(\nabla L)_{\mathbf{D}_y} \Leftarrow \frac{1}{|S|} \sum_i (\nabla L_i)_{\mathbf{D}_y}$
10:     $\mathbf{D}_x \Leftarrow \mathbf{D}_x - \nu_t (\nabla L)_{\mathbf{D}_x} / \|(\nabla L)_{\mathbf{D}_x}\|_F$
11:     $\mathbf{D}_y \Leftarrow \mathbf{D}_y - \nu_t (\nabla L)_{\mathbf{D}_y} / \|(\nabla L)_{\mathbf{D}_y}\|_F$
12:     Normalize atoms of $\mathbf{D}_x$ and $\mathbf{D}_y$
13:     $t \Leftarrow t + 1$
14: until convergence
```

4.3. Gradient Calculation

We follow the procedure put forward in [2] to calculate the gradient of our objective function (L) using the chain rule. The gradients for sample i can be calculated as follows

$$(\nabla L_i)_{\mathbf{D}_x} = \frac{\partial L_i}{\partial \mathbf{D}_x} + \frac{\partial L_i}{\partial \mathbf{z}_i^x} \cdot \frac{\partial \mathbf{z}_i^x}{\partial \mathbf{D}_x}, \qquad (4)$$

$$(\nabla L_i)_{\mathbf{D}_y} = \frac{\partial L_i}{\partial \mathbf{D}_y} + \frac{\partial L_i}{\partial \mathbf{z}_i^y} \cdot \frac{\partial \mathbf{z}_i^y}{\partial \mathbf{D}_y}. \qquad (5)$$

These equations are valid provided that the partial derivatives exist. The first two partial derivatives in each equation are taken on continuous functions and therefore exist everywhere. They can be calculated as follows.

$$\begin{aligned} &\frac{\partial L_i}{\partial \mathbf{D}_x} = 2(\mathbf{D}_x\mathbf{z}_i^x - \mathbf{x}_i)\mathbf{z}_i^{xT}, \quad \frac{\partial L_i}{\partial \mathbf{D}_y} = 2(\mathbf{D}_y\mathbf{z}_i^y - \mathbf{y}_i)\mathbf{z}_i^{yT}, \\ &\frac{\partial L_i}{\partial \mathbf{z}_i^y} = 2{\mathbf{D}_x}^T(\mathbf{D}_x\mathbf{z}_i^x - \mathbf{x}_i) + 2\gamma(\mathbf{z}_i^x - \mathbf{z}_i^y), \\ &\frac{\partial L_i}{\partial \mathbf{z}_i^y} = 2{\mathbf{D}_y}^T(\mathbf{D}_y\mathbf{z}_i^y - \mathbf{y}_i) - 2\gamma(\mathbf{z}_i^x - \mathbf{z}_i^y). \end{aligned}$$

More care must be taken when calculating the third partial derivative in each equation because they are not guaranteed to exist. In [2] it is shown that, given a reasonable prior on the input signals, they will exist with a very high probability. Even in the case when they don't, it is argued that the proposed calculations will still provide a feasible descent direction. Let Λ_x and Λ_y signify the sets of nonzero sparse coefficients for the gallery and probe domains, respectively. Therefore, Λ_x^c and Λ_y^c signify the sets of sparse coefficients equal to zero for the gallery and probe domains, respectively. The partial derivatives can be calculated as follows

$$\begin{aligned} &\frac{\partial \mathbf{z}_{i\Lambda_x}^x}{\partial \mathbf{D}_x} = (\mathbf{D}_{x\Lambda_x}^T \mathbf{D}_{x\Lambda_x})^{-1} \left(\frac{\partial \mathbf{D}_{x\Lambda_x}^T \mathbf{x}}{\partial \mathbf{D}_{x\Lambda_x}} - \frac{\partial \mathbf{D}_{x\Lambda_x}^T \mathbf{D}_{x\Lambda_x}}{\partial \mathbf{D}_{x\Lambda_x}} \mathbf{z}_{\Lambda_x} \right) \\ &\frac{\partial \mathbf{z}_{i\Lambda_y}^y}{\partial \mathbf{D}_y} = (\mathbf{D}_{y\Lambda_y}^T \mathbf{D}_{y\Lambda_y})^{-1} \left(\frac{\partial \mathbf{D}_{y\Lambda_y}^T \mathbf{y}}{\partial \mathbf{D}_{y\Lambda_y}} - \frac{\partial \mathbf{D}_{y\Lambda_y}^T \mathbf{D}_{y\Lambda_y}}{\partial \mathbf{D}_{y\Lambda_y}} \mathbf{z}_{\Lambda_y} \right) \\ &\frac{\partial \mathbf{z}_{i\Lambda_x^c}^x}{\partial \mathbf{D}_x} = 0 \qquad \frac{\partial \mathbf{z}_{i\Lambda_y^c}^y}{\partial \mathbf{D}_y} = 0. \end{aligned}$$

Once the gradients (4) and (5) are obtained, the coupled dictionaries are updated as shown in lines 10 and 11 of Algorithm 1. In this paper we set $\nu_t = \frac{.1}{\sqrt{t}}$ and $|S| = 1$.

5. EXPERIMENTS

We illustrate the effectiveness of our method on a thermal infrared face dataset that was created by the Wright State Research Institute at Wright State University. We used a subset of the dataset containing 65 subjects, each of which has between 14 and 33 samples. Each sample consists of a visible image and a Mid-wave IR image taken concurrently.

We extract features from the face images in the same manner as [1]. We first preprocess the images by filtering with a difference of Gaussian (DoG) filter and aligning based on fiducial points. We then partition each image into 143 overlapping 64 by 64 pixel blocks. Next, we compute HOG features of the blocks and reduce their dimension with principle component analysis (PCA). Finally we use the first 50 PCA components as input signals to our dictionary learning algorithm.

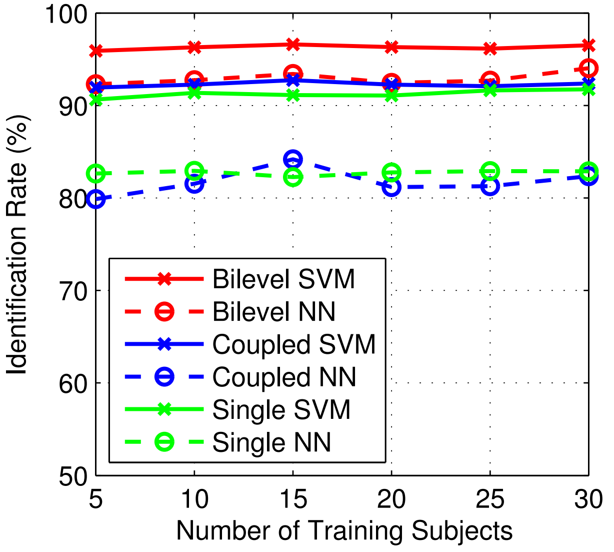

Fig. 1. Thermal face recognition performance on a 35 subject test set for a varying number of training subjects. We fix the dictionary size to 125 atoms for these experiments.

We test dictionaries learned by three different optimization methods: single, coupled and bilevel. The single dictionary method solves (1) with input signals from both domains. The coupled dictionary method solves (3) and the bilevel dictionary method solves (2). We pair each dictionary learning algorithm with two classification algorithms: linear Support Vector Machine (SVM) [14] and Nearest Neighbor (NN). We train the SVMs in a one versus all manner for each class. As previously mentioned, we partition the subjects in the dataset into two disjoint groups. The first group is used to train the dictionaries while the second group is used to evaluate the performance. For all results, we repeat the partitioning and experiment processes ten times and report the average performance. The classifiers are trained on sparse coefficients generated from the visible light images using the visible domain dictionary and tested on the sparse coefficients generated from the infrared images using the infrared domain dictionary.

Figure 1 shows the performance of our method based on the number of training subjects used. While the single dictionary performs reasonably well, the superior results of the bilevel dictionaries support our view that a hand picked feature extraction does not optimally bridge the thermal and visible domains. It is also important to note that the bilevel dictionaries outperform the coupled dictionaries. This shows the importance of the bilevel constraints. For the rest of the experiments, we fix the number of training and testing subjects to 10 and 55 respectively. Figure 2 and Figure 3 show the performance for varying values of γ and K. The identification rate does not vary much with γ, thus it does not need to be finely tuned. The same can be said for K with the SVM classifier. The NN classifier, on the other hand, is affected by changes in K. This is to be expected as NN tends to perform worse as the feature dimensionality increases [15].

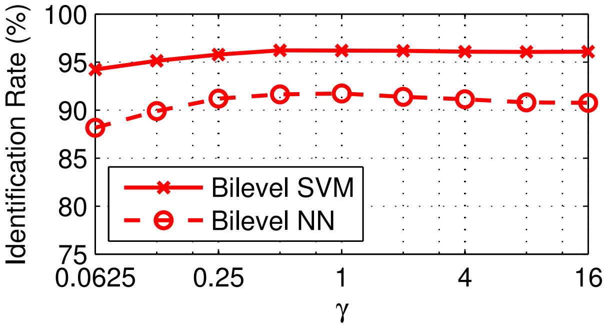

Fig. 2. Thermal face recognition performance for varying values of γ.

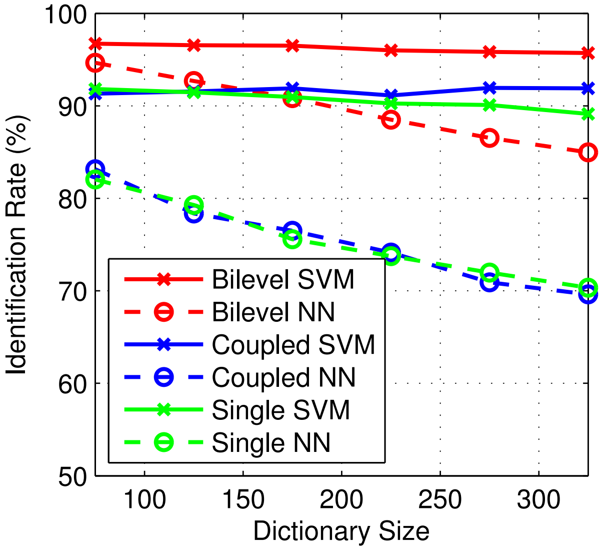

Fig. 3. Thermal face recognition performance for varying dictionary sizes.

6. CONCLUSION

We have presented a new approach to the problem of thermal to visible face recognition. We proposed the use of coupled dictionaries to bridge the gap between the thermal and visible domains. We then presented results of experiments demonstrating the efficacy of our approach.

7. REFERENCES

[1] Jonghyun Choi, Shuowen Hu, S. Susan Young, and Larry S. Davis, "Thermal to visible face recognition," 2012, vol. 8371, pp. 83711L–83711L–10.

[2] Jianchao Yang, Zhaowen Wang, Zhe Lin, Xianbiao Shu, and Thomas Huang, "Bilevel sparse coding for coupled feature spaces," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2360–2367.

[3] John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma, "Robust face recognition via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 210–227, 2009.

[4] M. Aharon, M. Elad, and A. Bruckstein, "k -svd: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, 2006.

[5] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma, "Image super-resolution via sparse representation," *Image Processing, IEEE Transactions on*, vol. 19, no. 11, pp. 2861–2873, 2010.

[6] Jianchao Yang, Zhaowen Wang, Zhe Lin, Scott Cohen, and Thomas Huang, "Coupled dictionary training for image super-resolution," *Image Processing, IEEE Transactions on*, vol. 21, no. 8, pp. 3467–3478, 2012.

[7] Shenlong Wang, Lei Zhang, Yan Liang, and Quan Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2216–2223.

[8] Kui Jia, Xiaogang Wang, and Xiaoou Tang, "Image transformation based on learning dictionaries across image spaces," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 2, pp. 367–380, 2013.

[9] Li He, Hairong Qi, and R. Zaretzki, "Beta process joint dictionary learning for coupled feature spaces with application to single image super-resolution," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, June 2013, pp. 345–352.

[10] Jingjing Zheng, Zhuolin Jiang, P Jonathon Phillips, and Rama Chellappa, "Cross-view action recognition via a transferable dictionary pair.," in *BMVC*, 2012, pp. 1–11.

[11] Thirimachos Bourlai, Arun Ross, Cunjian Chen, and Lawrence Hornak, "A study on using mid-wave infrared images for face recognition," in *SPIE Defense, Security, and Sensing*. International Society for Optics and Photonics, 2012, pp. 83711K–83711K.

[12] Brendan F. Klare and Anil K. Jain, "Heterogeneous face recognition using kernel prototype similarities," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 6, pp. 1410–1422, 2013.

[13] R.S. Ghiass, O. Arandjelovic, H. Bendada, and X. Maldague, "Infrared face recognition: A literature review," in *Neural Networks (IJCNN), The 2013 International Joint Conference on*, Aug 2013, pp. 1–10.

[14] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[15] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft, "When is nearest neighbor meaningful?," in *Database Theory ICDT99*, Catriel Beeri and Peter Buneman, Eds., vol. 1540 of *Lecture Notes in Computer Science*, pp. 217–235. Springer Berlin Heidelberg, 1999.