Facial Pose Estimation via Dense and Sparse Representation

Hui Yu, Honghai Liu, *Senior Member, IEEE*
University of Portsmouth, Portsmouth, PO1 2DJ, UK
{hui.yu; honghai.liu}@port.ac.uk

***Abstract*—Facial pose estimation is an important part for facial analysis such as face and facial expression recognition. In most existing methods, facial features are essential for facial pose estimation. However, occluded key features and uncontrolled illumination of face images make the facial feature detection vulnerable. In this paper, we propose methods for facial pose estimation via dense reconstruction and sparse representation but avoid localizing facial features. The Sparse Representation Classifier (SRC) method has achieved successful results in face recognition. In this paper, we explore SRC in pose estimation. Sparse representation learns a dictionary of base functions, so each input pose can be approximated by a linear combination of just a sparse subset of the bases. The experiment conducted on the CMU MultiPIE face database has shown the effectiveness of the proposed method.**

Keywords—linear regression; pose analysis; human face; 3D face

I. INTRODUCTION

The wide applications of facial pose estimation in human-computer interaction, computer vision and mobile policing etc. have attracted attention in recent years.

Extensive effort has been done to estimate head poses through detecting facial features and locating landmarks on faces [9, 10].

The existing methods can be categorized into two categories: feature-based method and holistic method. The feature-based methods usually depend on localizing key facial features such as mouth corners, eye corners and nose tip etc. [2]. This method assumes that the configuration of key landmarks does not change significantly with various facial expressions for the same identity. It can achieve a fast estimation result. However, due to uncertainty of the quality of face images under uncontrolled environments, it is still challenging to detect facial features and landmarks. This ambiguity in facial feature detection can lead to poor performance in pose estimation. The holistic method utilizes the facial appearance for pose estimation [1, 13]. It does not need any feature points for pose estimation. Facial appearance carries rich information of a person such as identity, facial expression, illumination and poses etc. In some cases the variation of information such as illumination can cause bigger challenges for facial poses leading to pronounced estimation errors.

In this paper, we propose to estimate facial poses using both dense reconstruction and Sparse Representation. The proposed method is based on the assumption that pose space cannot be approximated by a combination of a pose subspace.

II. DENSE AND SPARSE REPRESENTATION

A. Sparse Representation

Sparse Representation (SR) has shown strong performance in applications of computer vision, especially face recognition [3, 4]. It learns a dictionary of basis functions, so each input signal can be approximated by a linear combination of just a sparse subset of the bases. The test signal can be formed as a linear combination of the training samples which is the category the test signal is supposed to fall in.

Our assumption is that the facial pose cannot be approximated by cross subspaces. Our earlier findings have supported this assumption [11]. In this paper, we seek a more compact solution to this problem with a sparse base combination for each facial pose.

The idea of Sparse Representation is to express a test signal through a sparse combination of bases in a sample dictionary. SR was developed based on the compressed sensing theory. It was originally explored in the signal pressing area for reconstructing a sparse signal based on a sparsity structure [5, 6]. By applying SR to face recognition, it is assumed that the face samples from the same identity class approximately span in a linear subspace. Any given query faces can be expressed as the linear combination of training samples of the same class in the training dictionary. For facial pose estimation, the given query pose can be linearly approximated by a combination of the same pose samples from an over-completed dictionary.

Suppose we have C classes of facial poses with each class being formed by n training samples of facial poses. These C classes of poses are formed as a over-completed dictionary denoted by a matrix X, $X = [x_1, ..., x_n] \in \Re^{p \times n}$, where $x_i \in \Re^p$ represents a training facial pose image arranged in a

This work was supported by EU seventh framework programme under grant agreement No. 611391, Development of Robot-Enhanced Therapy for Children with Autism Spectrum Disorders (DREAM).

column-wised vector. The test facial pose $y \in \Re^p$ is expected to be expressed as a sparse representation of samples of the matrix X, such that each x_i contains k (k<<n) or fewer nonzero elements. This can be described as the following optimization problem:

$$\hat{w} = \arg\min_w \| w \|_0 \text{ subject to } y = Xw \quad (1)$$

where $w \in \Re^K$ is vector of weights representing the contribution of each facial pose in X and $\| w \|_0$ is l_0 -norm.

Equation (1) is under-determined, so it is NP-hard to find the solution to this sparse representation problem [13]. Thus, an alternative solution is to solve the l_1-minimization problem. The l_1 -minimization is more robust to outliers, so it can efficiently reconstruct the sparse signal. The objective function can be written as follows:

$$\hat{w} = \arg\min_w \|y - Xw\|_2 + \lambda \|w\|_1 \quad (2)$$

Where λ is a parameter which is used to regularize the residual and the sparsity.

Each element of the learnt weight vector $\hat{w}$ is associated with C class labels of the training samples. Thus, $\hat{w}$ can be expressed as $\hat{w} = [\hat{w}_1, \hat{w}_2, ..., \hat{w}_C]$. Specifically, those nonzero entries of $\hat{w}_i$ represents a subset of the weighting vector for the i^{th} class corresponding to training samples X_i . With these weighting parameters $\hat{w}_i$, the pose can be reconstructed as follows:

$$\hat{y}_i = X_i \hat{w}_i \quad (3)$$

The query pose y can be classified according to the minimal error between y and the reconstructed $\hat{y}$:

$$e_i = \frac{\| y - \hat{y}_i \|_2}{N} \quad (4)$$

where N is the total pixel number of the face image.

B. Dense Reconstuction

The dense reconstruction of a given facial pose is actually a linear regression model for the training samples. Various methods have been proposed in reconstruction of faces or facial expressions for different applications [8, 12]. We apply the linear regression method in this paper. We expect the given facial pose can be approximately reconstructed through linear combination of examples with the same or similar poses. Thus, the dictionary matrix is formed by m vectorized pose images represented by $X = [X_1, X_2, ...X_m]$, $X \in \Re^{p \times m}$. To reduce the dimension of the matrix, Principal Component Analysis (PCA) is applied to X to obtain a compact representation:

$D = [D_1, D_2, ..., D_q]$, $D \in \Re^{l \times q}$, where D is the eigenvectors corresponding to the first q eigenvalues. Thus, the reconstructed facial pose based on the training dictionary can be calculated as follows:

$$\hat{y} = DW + \overline{X} \quad (5)$$

where W is the reconstruction coefficient vector: $W = D^T (y - \overline{X})$ and $\overline{X}$ is the mean facial pose of the dictionary D.

The reconstruction error between the query pose image and the reconstructed pose can be calculated using (4).

III. EXPERIMENT

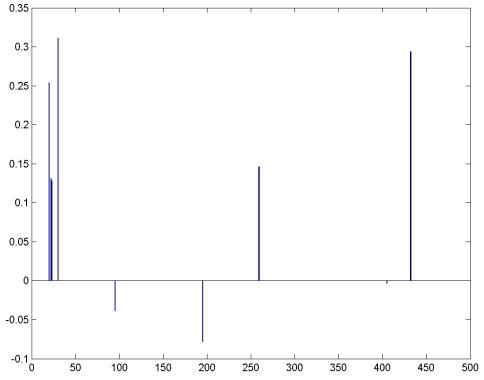
.

Fig. 1. Examples of 13 viewpoints plus 2 additional viewpoints simulating typical surveillance camera views.

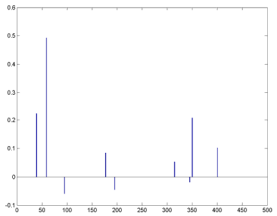
Fig. 2. An example of cropped face image with the size of 128×128 pixels

We test the proposed methods on the CMU MultiPIE face database [7]. The CMU MultiPIE database consists of 337 subjects with each subject's face taken from 13 viewpoints plus 2 additional viewpoints simulating typical surveillance camera views. Fig. 1 shows examples of those 15 poses of face images.

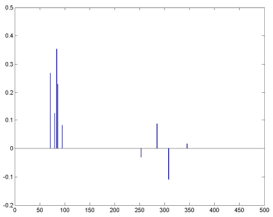
The experiment was conducted using 30 training facial pose images for each pose. There are 15 poses in total. Thus the total dimension of the training samples is 450. Each raw face image is cropped to size 128×128 as shown in Fig. 2. The weight vector with sparsity is achieved through minimizing (2). Fig. 3 shows examples of sparse representation of weight vectors for 15 poses (poses -90, -75, -60, -45R, -45, -30, -15, 0, 15, 30, 45, 45R, 60, 75, 90) of identity one. Two poses (45R and -45R) were captured by cameras located above the subject, simulating a typical surveillance camera view. In the training samples, there are 30 training poses for each pose arranged in 450 columns in the training sample matrix X. Thus, every 30 columns in X represent one pose out 15 poses. As shown in Fig. 3 (a), the values from 0 to 30 in X-axis represent the weight vector of pose one. The sparse nonzero values within 30 in X-axis illustrates the strong weight for pose one (-90). There are two values in around 260 and 435, but they have weak affect compared with those within 30. Thus, the query pose is expected to belong to pose one in this case. The error calculated using (4) verifies the prediction. Figures (a) - (d) in Fig. 4 illustrate the reconstruction errors of the first four poses respectively: -90, -75, -60, -45R. The smallest error value in each figure means that the query pose is classified to that pose indicating by the value on X-axis. It shows that the errors for classification have clear dissimilarity indicating the effectiveness of the method.



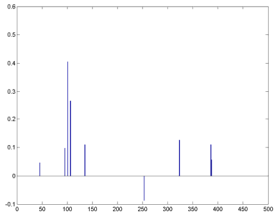
(a)



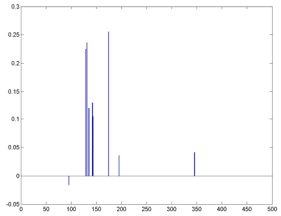
(b)



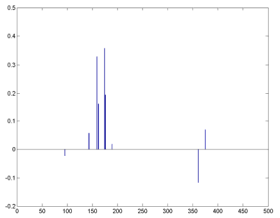
(c)



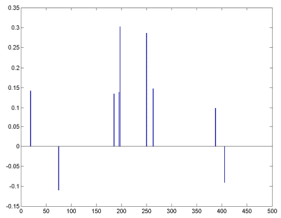
(d)



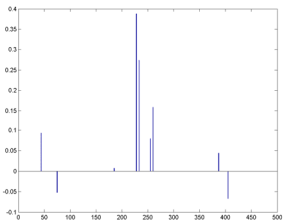
(e)



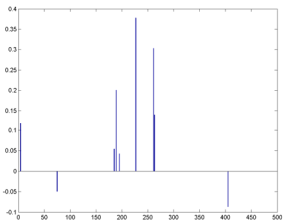
(f)



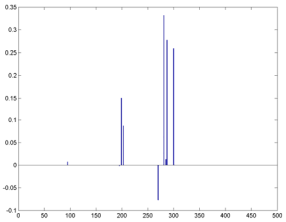
(g)



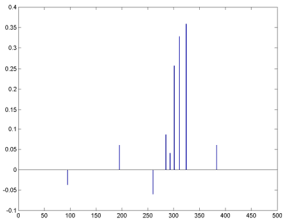
(h)



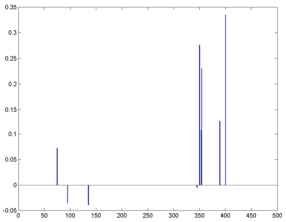
(i)



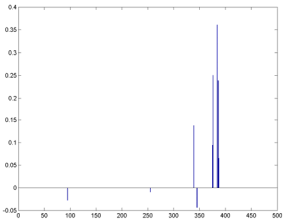
(j)



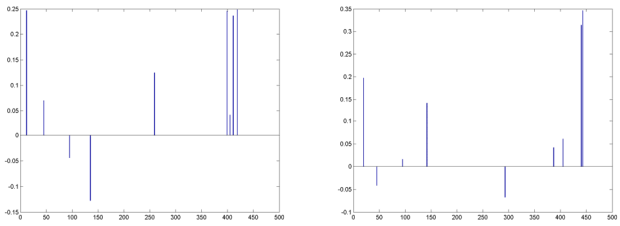
(k)



(l)

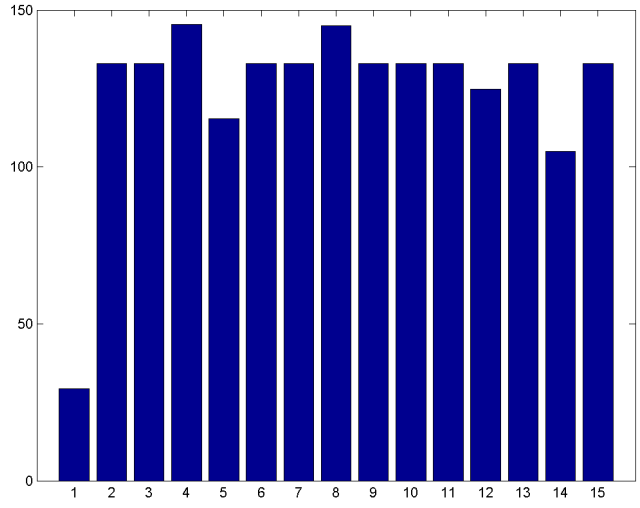


(m)

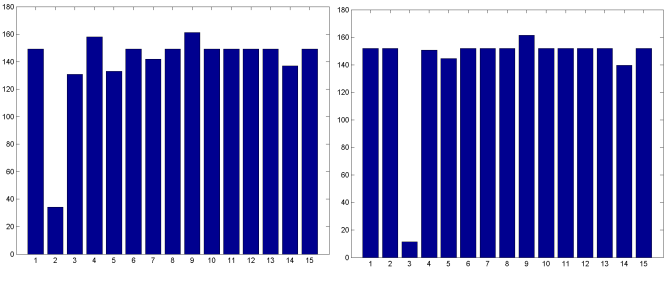


(o) (p)

Fig.3. Visualization of the sparse weight vector W for the query pose one. From (a) to (p) are sparse representation for 15 poses -90, -75, -60, -45R, -45, -30, -15, 0, 15, 30, 45, 45R, 60, 75, 90 respectively. Two poses (45R and -45R) were captured by cameras located above the subject, simulating a typical surveillance camera view. The X-axis represents the dimension of training samples. Y-axis indicates the value for the weights of their corresponding training samples.

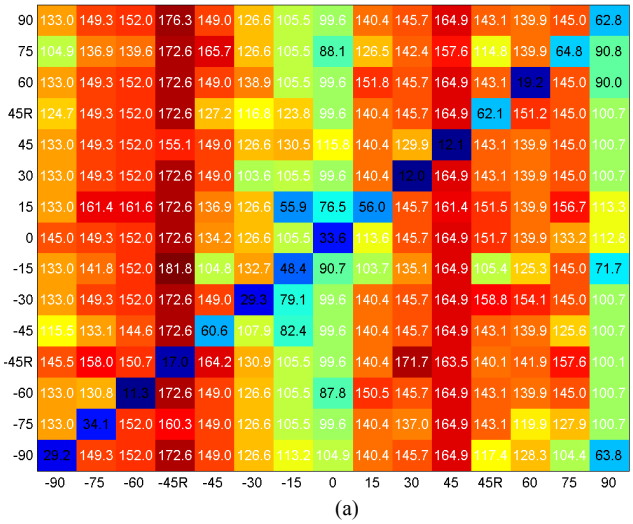


(a)

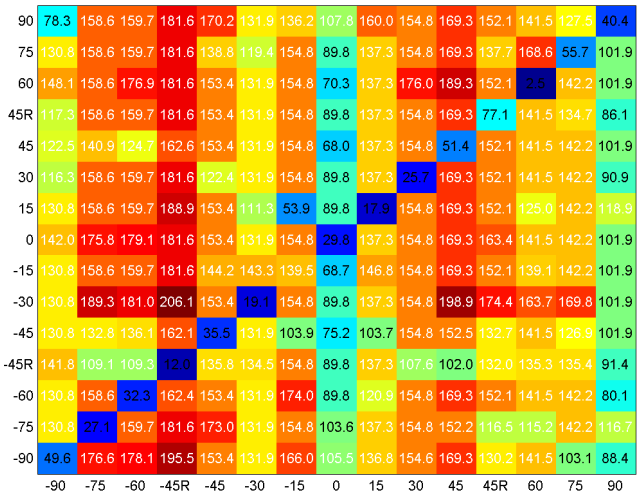


(b) (c)

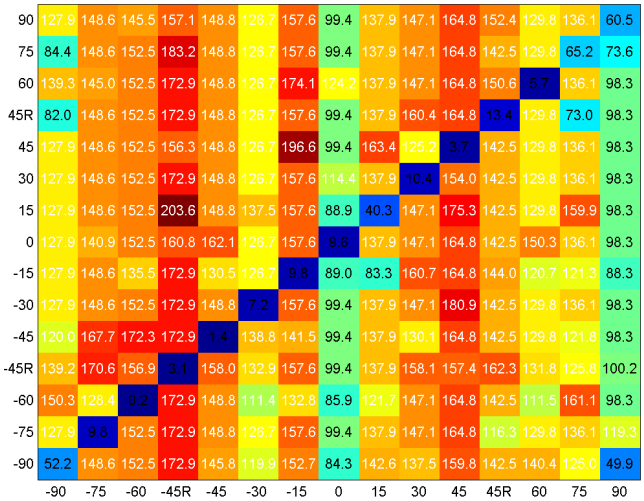
Fig. 4. Examples of reconstruction errors using the Sparse Representation Method. Figures from (a) to (c) illustrate errors of the first 3 query poses of identity one with pose being -90, -75 and -60 respectively. The smallest error in each figure means that the query pose has the same pose with this pose category.



(a)



(b)



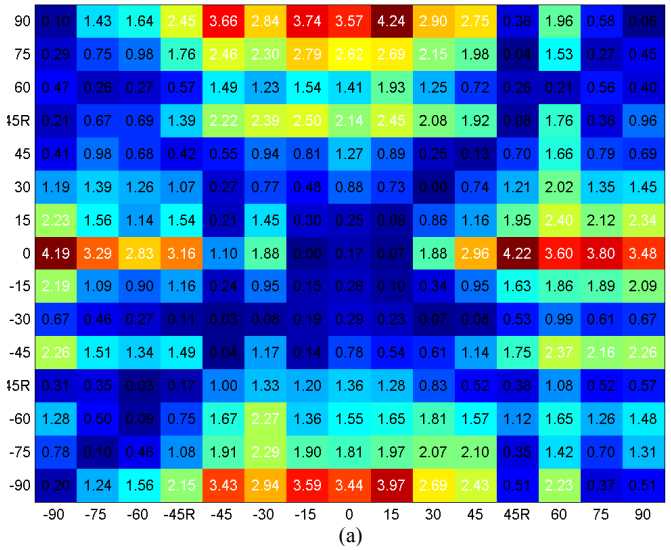
(c)

Fig. 5. Examples of dissimilarity matrix of pose estimation. Figures from (a) to (c) demonstrate the reconstruction errors of 15 poses for 3 identities.

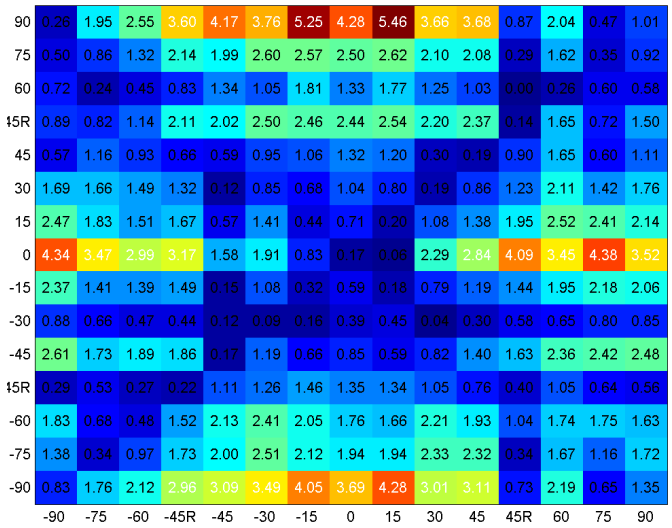
We have done experiments using the dense reconstruction method on the same database. We tested all 15 poses with each pose containing 50 identities. Thus the size of the training sample is $p \times 50$, where p is the pixel number of the image.

Fig. 5 demonstrates dissimilarity matrix of reconstruction errors of 15 poses for 3 identities using the Sparse Representation method. We can see the query poses can be clearly reconstructed by the same pose sub-dictionary with much lower errors than the others. This can be reflected by the much smaller error values along the diagonal of those dissimilarity matrices in the figures.

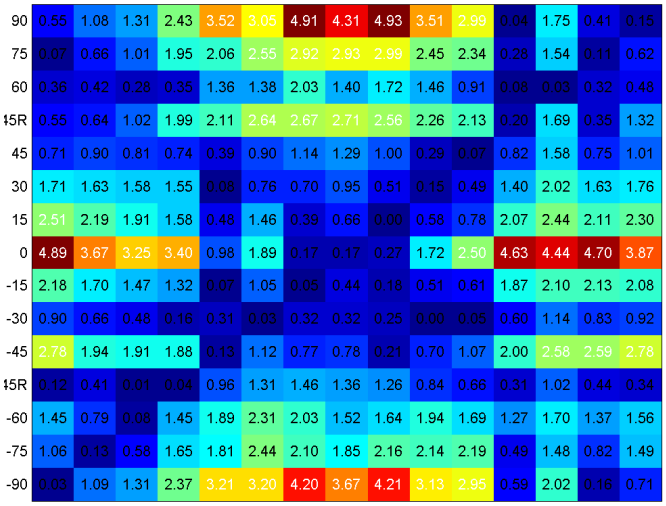
The dissimilarity matrix in Fig. 6 shows that dense reconstruction cannot guarantee a reliable reconstruction of the query pose, though it achieved success for some poses. This is reflected by the errors along the diagonal of each dissimilarity matrix, which are not remarkably smaller than the rest. Compared with dense reconstruction classification, Sparse Representation can achieve a better performance in pose estimation.



(a)



(b)



(c)

Fig. 6. Examples of dissimilarity matrix of pose estimation from dense reconstruction. Fig. (a) to (c) demonstrate the reconstruction errors of 15 poses for three query identities respectively.

IV. CONCLUSION

In this paper, we have proposed to estimate facial poses through dense reconstruction and Sparse Representation. The dense reconstruction method reconstructs the facial pose based on the linear combination of the pose dictionary. It can approximately reconstruct the query pose but less reliability. It is noticed that pose with the same angle but opposite view points have similar reconstruction errors in some cases. The experiments have demonstrated that Sparse Representation has better performance than the dense reconstruction in facial pose estimation.

REFERENCES

[1] Y. Ma, Y. Konishi, K. Kinoshita, S. Lao, and M. Kawade. Sparse bayesian regression for head pose estimation. InProc. ICPR, pages 507–510, 2006.

[2] Jian-Gang Wang, Eric Sung, EM enhancement of 3D head pose estimated by point at infinity Image and Vision Computing Volume 25 Issue 12, December, 2007 Pages 1864-1874

[3] Wright J, Ma Y, Mairal J, Sapiro G, Huang TS, et al. Sparse Representation for Computer Vision and Pattern Recognition. Proceedings of the IEEE 98: 1031–1044, 2010

[4] Wagner A, Wright J, Ganesh A, Zihan Z, Mobahi H, et al. Toward a Practical Face Recognition System: Robust Alignment and Illumination by Sparse Representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34: 372–386. 2012

[5] Donoho DL, Compressed sensing. IEEE Transactions on Information Theory 52: 1289–1306.

[6] Candes EJ, Romberg J, Tao T (2006) Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. IEEE Transactions on Information Theory 52: 489–509.

[7] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, & S. Baker (2008). Multi-PIE. Proceedings of the Eighth IEEE International Conference on Automatic Face and Gesture Recognition.

[8] Yu, H., Liu, H. (2014) Regression-Based Facial Expression Optimization, , IEEE Transactions on Human-Machine Systems, Vol 44 Issue 3, pp 386-394.

[9] E. Murphy-Chutorian and M. Trivedi. Head pose estimation in computer vision: A survey. IEEE Trans. Pattern Anal. Mach. Intell, vol.31, no.4, pp.607,626, April 2009.

[10] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld. Face recognition: A literature survey. ACM Computing Surveys, Volume 35 Issue 4, December 2003 Pages 399-458.

[11] Yu, H., Liu, H. Linear Regression for Head Pose Analysis, , IEEE World Congress on Computational Intelligence, Beijing, China, 2014;.

[12] Naseem I, Togneri R, Bennamoun M. Linear regression for face recognition, IEEE Trans Pattern Anal Mach Intell. 2010 Nov;32(11):2106-12. doi: 10.1109/TPAMI.2010.128.

[13] Y. Li, S. Gong, J. Sherrah, and H. Liddell. Support vector machine based multi-view face detection and recognition. Image and Vision Computing, 22(5):413–427, 2004.