# Video-to-video Pose and Expression Invariant Face Recognition using Volumetric Directional Pattern

Almabrok E. Essa and Vijayan K. Asari
*Department of Electrical and Computer Engineering*
*University of Dayton, 300 College Park, Dayton, OH, 45469, U.S.A.*

Keywords: Volumetric Directional Pattern, Key Frame Extraction, Video to Video Face Recognition.

Abstract: Face recognition in video has attracted attention as a cryptic method of human identification in surveillance systems. In this paper, we propose an end-to-end video face recognition system, addressing a difficult problem of identifying human faces in video due to the presence of large variations in facial pose and expression, and poor video resolution. The proposed descriptor, named Volumetric Directional Pattern (VDP), is an oriented and multi-scale volumetric descriptor that is able to extract and fuse the information of multi frames, temporal (dynamic) information, and multiple poses and expressions of faces in input video to produce feature vectors, which are used to match with all the videos in the database. To make the approach computationally simple and easy to extend, key-frame extraction method is employed. Therefore, only the frames which contain important information of the video can be used for further processing instead of analysing all the frames in the video. The performance evaluation of the proposed VDP algorithm is conducted on a publicly available database (YouTube celebrities' dataset) and observed promising recognition rates.

## 1 INTRODUCTION

Face detection and recognition has received a great deal of attention for the past three decades and become one of the most popular research areas in computer vision and pattern recognition. Face recognition has spread in several applications such as biometric systems, access control and information security systems, content-based video retrieval systems, and more generally image understanding. Only recently researchers' interest spread into the domain of video, where the problem becomes more challenging due to the pose variations, different facial expressions, illumination changes, occlusions and so on. However, it also has the benefit of providing many samples of the same person, thus providing the opportunity to convert many weak examples into a strong prediction of the identity. Due to its importance, many researchers have focused on video based face recognition with several approaches proposed as in (Shakhnarovich et. al., 2002), (Liu and Chen, 2003), (Lee et. al., 2003), (Aggarwal et. al., 2004), (Arandjelovi´c and Cipolla, 2009), (Nishiyama et. al., 2005) and (Li et. al., 2005).

Face recognition can generally be categorized into one of three scenarios based on the characteristics of the images to be matched, such as still-to-still image, video-to-still image, and video-to-video face recognition. Unlike still faces videos contain plentiful information than a single image like spatiotemporal information. Face recognition in videos can be more robust and stable by fusing information of multi frames, temporal information and multi poses of faces in videos make it possible to explore shape information of the face and combined into the framework of face recognition (Suneetha, 2014). The video-based recognition has more advantages than the image-based recognition. Firstly, the temporal information of faces can be utilized to facilitate the recognition task. Secondly, more effective representations, such as a 3D face model or super-resolution images, can be obtained from the video sequence and can be used to improve recognition results. Finally, video based recognition allows learning or updating the subject model over time to improve recognition results for future frames. So video based face recognition is also a very challenging problem, which suffers from following nuisance factors such as low quality facial images, scale variations, illumination changes, pose variations, motion blur, and occlusions and so on (Zhang et. al, 2011) (Best-Rowden et. al., 2013).

The key of each face recognition system is the utilization of the feature extraction technique that


Asari V. and Essa A..
Video-to-video Pose and Expression Invariant Face Recognition using Volumetric Directional Pattern.
DOI: 10.5220/0005353604980503
In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications* (VISAPP-2015), pages 498-503
ISBN: 978-989-758-090-1

must be able to extract features from the face image that is distinct and stable under different conditions during the image acquisition process. The texture of objects in digital images is an important property utilized in many computer vision and image analysis applications such as face recognition, object classification, and segmentation. After obtaining the image of a face, the next step of human face identification is to extract and describe salient features from facial images, in the context of feature description and representation applications, there are two common types of techniques; the first one is subspace based holistic feature which suffer during illumination variation and alignment error. The second one is local appearance feature. Among the most successful local face appearance representations are local patterns which are basically fine-scale descriptors that capture small texture details.

The rest of this paper is organized as follows. Section II illustrates the video to video face recognition methodology, section III presents the proposed VDP approach, section IV discusses the key frame extraction technique, section V presents the experimental results, and section VI concludes the paper.

## 2 METHODOLOGY

The main goal of the volumetric directional pattern is extracting and fusing the temporal information (dynamic features) from three consecutive frames which are distinct under multi poses and facial expression variations. Many of the computer vision applications employ the texture analysis algorithms. One of the most high performing texture algorithms is based on the concept of local pattern descriptor which describes the relationship of pixels to their local neighbourhood. To make the approach easy to extend and computationally simple, the key-frames based method is performed. Instead of analysing all the frames in the video, only the frames which contain more information of the video can be used for feature extraction processing. The key frames are obtained by computing the edge difference between the consecutive frames and those frames exceeding a predefined threshold are considered as key frames. The overall processing steps in the proposed technique is presented in figure 1.

With a video as input and a gallery of videos, we perform face recognition process throughout the whole video clip. Firstly, we detect and track faces using Viola and Jone's face detector. Then for each frame we extract and combine the dynamic features

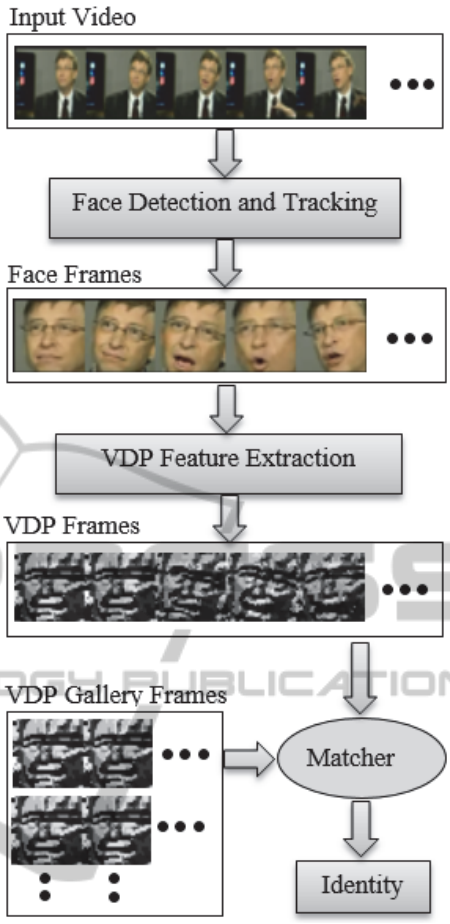


Figure 1: Video-to-video face recognition pipeline.

of its two neighbouring frames using our new algorithm. Similar procedure is flowed for the gallery videos. Finally, we compare the encoded VDP-feature histogram from each frame with all other candidate's VDP-feature vectors of all gallery video frames using Chi-Square dissimilarity measure.

## 3 VOLUMETRIC DIRECTIONAL PATTERN (VDP)

The extraction of distinct and stable features is the most fundamental and important problem in pattern recognition. The first step after obtaining the image of a face is to extract and describe salient features. There are two common types of features; the first one is a subspace based holistic feature (geometric features). The second one is local appearance feature. Geometric features can be represented by the shapes and location of facial components such as eyes, nose, mouth etc. Due to the uncontrolled environments

during the image acquisition process, these representations suffer to obtain the facial features. The appearance based features present the appearance changes of the facial skin texture, such as wrinkles and furrows (Suneetha, 2014). Among the most successful local face appearance representations are local patterns which are basically fine-scale descriptors that capture small texture details.

Volumetric Directional Pattern (VDP) represents the local appearance features. VDP is a gray-scale pattern that characterizes and fuses the temporal structure (dynamic information) of three consecutive frames. VDP computes the edge response values in different directions at each pixel position, and uses the relative strength magnitude to encode the image texture. The VDP feature extraction techniques uses the concept of local directional pattern extraction method presented in (Jabid et. al., 2010), (Kim et. al., 2013). VDP is a twenty four bit binary code assigned to each pixel of an input frame. This can be calculated by comparing the relative edge response value of a particular pixel from three consecutive frames in different directions by using Kirsch masks in eight different orientations $(M_0 - M_7)$. The masks are centred on its own position for one frame and the corresponding positions of the other two frames. These masks are shown in figure 2.

$$M_0 = \begin{bmatrix} -3 & -3 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & 5 \end{bmatrix} \quad M_1 = \begin{bmatrix} -3 & 5 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & -3 \end{bmatrix} \quad M_2 = \begin{bmatrix} 5 & 5 & 5 \\ -3 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix} \quad M_3 = \begin{bmatrix} 5 & 5 & -3 \\ 5 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix}$$

$$M_4 = \begin{bmatrix} 5 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & -3 & -3 \end{bmatrix} \quad M_5 = \begin{bmatrix} -3 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & 5 & -3 \end{bmatrix} \quad M_6 = \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & -3 \\ 5 & 5 & 5 \end{bmatrix} \quad M_7 = \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & 5 \\ -3 & 5 & 5 \end{bmatrix}$$

Figure 2: Kirsch edge masks in all eight directions.

Given a central pixel in the current frame (middle frame) of three consecutive frames, the eight different directional edge response values are $c_i (i = 8, 9, \ldots, 15)$ and that used to create an eight bit binary number. This can describe the edge response pattern of each pixel in the current frame. The eight different edge response values $p_i (i = 16, 17, \ldots, 23)$ and $n_i (i = 0, 1, \ldots, 7)$ that are used to create an eight bit binary number. Each number can describe the edge response pattern of the respective pixel in the previous frame and next frame respectively. Finally, we concatenate all of these spatiotemporal directional response patterns of a particular pixel to get a twenty four bit binary pattern. These edge responses and the corresponding bit positions are shown in figure 3.

| | | |
|---|---|---|
| $p_{19}$ | $p_{18}$ | $p_{17}$ |
| $p_{20}$ | $X$ | $p_{16}$ |
| $p_{21}$ | $p_{22}$ | $p_{23}$ |

| | | |
|---|---|---|
| $b_{19}$ | $b_{18}$ | $b_{17}$ |
| $b_{20}$ | $X$ | $b_{16}$ |
| $b_{21}$ | $b_{22}$ | $b_{23}$ |

(a)

| | | |
|---|---|---|
| $c_{11}$ | $c_{10}$ | $c_9$ |
| $c_{12}$ | $X$ | $c_8$ |
| $c_{13}$ | $c_{14}$ | $c_{15}$ |

| | | |
|---|---|---|
| $b_{11}$ | $b_{10}$ | $b_9$ |
| $b_{12}$ | $X$ | $b_8$ |
| $b_{13}$ | $b_{14}$ | $b_{15}$ |

(b)

| | | |
|---|---|---|
| $n_3$ | $n_2$ | $n_1$ |
| $n_4$ | $X$ | $n_0$ |
| $n_5$ | $n_6$ | $n_7$ |

| | | |
|---|---|---|
| $b_3$ | $b_2$ | $b_1$ |
| $b_4$ | $X$ | $b_0$ |
| $b_5$ | $b_6$ | $b_7$ |

(c)

Figure 3: Twenty four edge responses and VDP binary bit positions: a) For the previous frame, b) for the current frame, c) for the next frame.

$$p_i = \sum_{i=16}^{23} dot(I_{3\times3}, M_{i-16}) \quad (1)$$

$$c_i = \sum_{i=8}^{15} dot(I_{3\times3}, M_{i-8}) \quad (2)$$

$$n_i = \sum_{i=0}^{7} dot(I_{3\times3}, M_i) \quad (3)$$

Where $dot(\cdot)$ represents the dot product operation, $M_i$ is the mask, and $I_{3\times3}$ is $3 \times 3$ neighbors of the center pixel of each frame. $p_i, c_i$ and $n_i$ are the spatiotemporal directional response values for the previous, current, and next frames respectively.

In order to generate the VDP-feature, we need to know the most prominent temporal information indicating the dynamic features (DF) for all three consecutive frames and then set them to 1 and the rest of the 16 bits of VDP pattern are set to 0. Finally, the VDP code can be derived by:

$$VDP = \sum_{i=0}^{7} b_i(n_i - n_{DF}) \times 2^i + \sum_{i=8}^{15} b_i(c_i - c_{DF}) \times 2^i + \sum_{i=16}^{23} b_i(p_i - p_{DF}) \times 2^i, \quad b_i(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (4)$$

Where $p_{DF}, c_{DF}$ and $n_{DF}$ the $DF^{th}$ most prominent dynamic features for all three consecutive frames. Figure 4 shows an example of VDP code computation with $DF = 3$. We replace each particular pixel from

one frame with the result of combining its 3×3 neighbours with the other 3×3 corresponding neighbours from the previous and next frames.

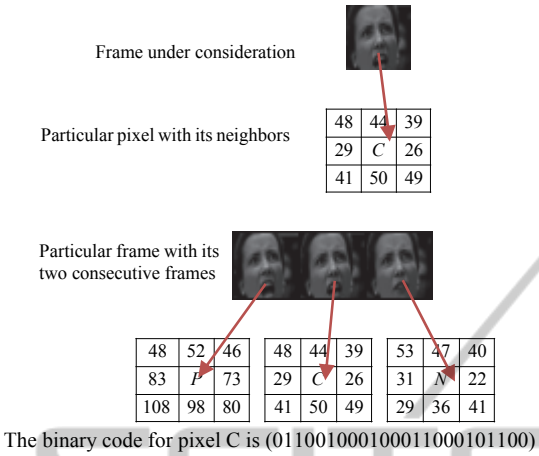


Figure 4: VDP code computation of single subject.

***The algorithm stages for VDP calculation:***

***Stage1:*** Detect faces from the video scene (we used Viola and Jone's face detector).

***Stage2:*** Select three consecutive frames each time and for each pixel and its corresponding pixels. Apply equation 1 for the previous frame's pixels, equation 2 for the current frame's pixels, and equation 3 for the next frame's pixels.

***Stage 3:*** Apply equation 4 by the time of finishing stage 2 for each pixel. Then substitute this value instead of that pixel.

***Stage 4:*** Normalize and extract the VDP-feature histogram for the matching process.

## 4 KEY FRAME EXTRACTION

An important step in video analysis and content based video information is key frame extraction which is an essential part in video summarization. Key frame is the frame which can represent the salient content and information of the video. The key frames extracted must summarize the characteristics of the video, and the image characteristics of a video can be tracked by all the key frames in time sequence (Khurana et. al., 2013), (Liu and Zhao, 2009).

It is not meaningful to analyze all the frames in the video if they do not contain important information. Therefore, we find and detect the frames which contain important information to use for further process. For the detection of key frames we have used the Canny edge detector to calculate the difference between two consecutive frames. Only when the difference exceeds a threshold, one of the consecutive frames is considered as a key frame.

The steps for key frame extraction from the video is as follows, and figure 5 shows the key frames and their VDP-features.

***Step1:*** Obtain the gray scale image for two consecutive frames.

***Step2:*** Use the Canny edge detector to find the edge difference between all consecutive frames.

***Step3:*** Compute the mean and standard deviation for all edge differences.

***Step4:*** Calculate the threshold using the formula: Threshold = mean + a × standard deviation where a is a constant.

***Step5:*** Find and store the key frames which exceed that threshold.

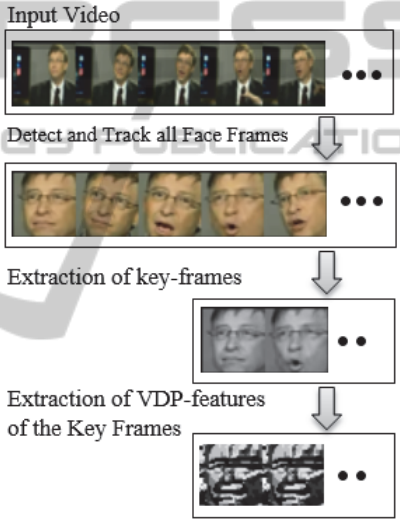


Figure 5: Extraction of key frames.

## 5 EXPERIMENTAL RESULTS

To evaluate the pose and expression variation robustness of the proposed method, we tested it on YouTube Celebrities dataset (Kim et. al., 2008). All the face images in this work were detected by using the Viola and Jone's face detector (Viola and Jones 2004). After key frame extraction process all the frames are resized to 64 × 64 and then extract the spatiotemporal information using the proposed VDP. When it comes to the face recognition process, we represent the face using a VDP-feature histogram. The objective is to compare the encoded feature vector from one frame with all other candidate's feature vector using Chi-Square dissimilarity

measure. This is a measure between two feature vectors, H1 and H2, of length $N$ and it is defined as:

$$\chi^2(H_1,H_2)=\sum_{i=1}^{N}\frac{(H_1(i)-H_2(i))^2}{H_1(i)+H_2(i)} \quad (5)$$

The corresponding face of the VDP-feature vector with the lowest measured value indicates the match found. Note that all the implementations are in the MATLAB environment using Intel(R) Xeon(R) desktop computer, 2.00GHz CPU with 12GB RAM.

## 5.1 YouTube Celebrities Dataset

It is a large-scale video dataset which contains 1910 video sequences of 47 different subjects (actors and politicians celebrities). The dataset is considered as one of the most challenging video databases due to the large illumination, pose, and expression variations as well as low resolution and motion blur, which can be seen in figure 6. We evaluated the proposed VDP on all the 47 celebrities, while the state-of-the-art methods evaluated on only a subset of the subjects (e.g. in (Yang et. al., 2013). They use the first 29 celebrities). For each subject, we select one video clip as the training data with a different one as the testing data from the same video scene. The experimental setup were organized as 2 parts. Part 1 uses all frames of the training/testing data that includes 15969 frames and part 2 uses only the key-frames of the training/testing data.

Figure 6: Some samples of YouTube celebrities' dataset.

### 5.1.1 Using All Frames

We tested our method for face recognition using all the frames from each video in the testing set which includes 7745 frames. We compared the features of each frame one by one with all the training set which includes 8224 frames. This has taken 0.43 s/frame for the whole process (detect faces, extract VDP-features, and matching process).

### 5.1.2 Using Key-frames

In this part we tested our method for face recognition using three kinds of key frame selections strategies depending on the edge difference between the frames. This is to show the performance and the effectiveness of the proposed method in terms of the accuracy and speed. In the first case, the number of frames in the testing set became 1063 frames and that in the training set became 1127 frames. This set is called K-F1. In the second case, the testing set became 2609 frames and the training set became 2712 frames. This is named as K-F2. Finally, the testing set is reduced to 3175 frames and the training set to 3319 frames.

This set is called K-F3. Figure 7 shows the computation time (time of matching process only) with respect to the number of frames and illustrates the effectiveness of choosing a lower number of frames by using the key-frame selection technique. The results including the recognition rate and the

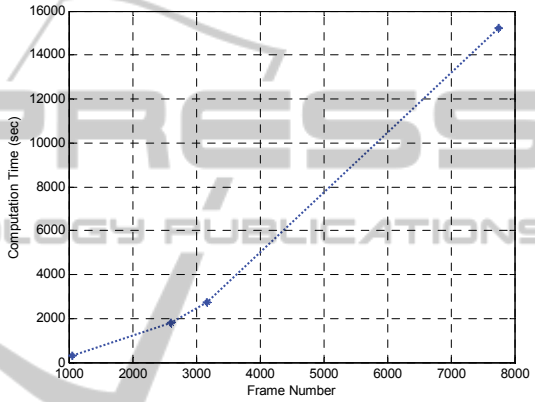


Figure 7: Computation time with respect to the number of frames in the training/testing set.

Table 1: Accuracy on YouTube celebrities dataset.

| Method | TP | FP | Rec. Rate | Speed |
|---|---|---|---|---|
| K-F1 | 824 f | 239 f | 77.5 % | 5.2 m/all |
| K-F2 | 2168f | 441 f | 83.09 % | 32.3 m/all |
| K-F3 | 2656f | 519 f | 83.65 % | 45.8 m/all |
| All-F | 7552f | 202 f | 97.39 % | 4.7 h/all |

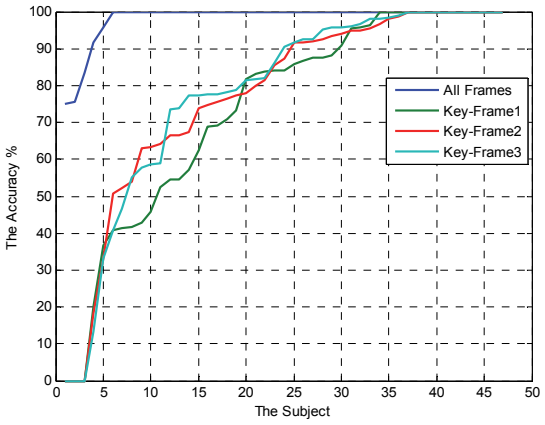


Figure 8: The accuracy for each subject for all cases.

computation time are summarized in Table 1. Figure 8 shows the accuracy of each subject with all 47 subjects in all cases. The plots are presented in an ascending order of recognition rate. It can be seen that in this dataset, around 42 out of 47 video sets of YouTube celebrities are recognizing really well. Three of the datasets are too bad for recognition even when we use all the frames for training.

# 6 CONCLUSIONS

In this paper, we proposed a volumetric directional pattern (VDP) approach for robust and fast video to video based face recognition. We developed a novel algorithm that has the ability to extract and fuse the temporal information for the analysis of facial dynamic changes. By using two video sequences of the same video scene per subject, we showed that our method could achieve higher identification accuracy than the state-of-the-art methods. In this paper we also presented the effect of key frame technique in terms of accuracy and speed.

# REFERENCES


G. Shakhnarovich, J. Fisher, and T. Darrell, 2002. Face recognition from long-term observations. Computer Vision ECCV.

X. Liu and T. Chen, 2003. Video-based face recognition using adaptive hidden markov models. In Computer Vision and Pattern Recognition. IEEE Computer Society Conference.

K. C. Lee, J. Ho, M.H. Yang and D. Kriegman, 2003. Video-based face recognition using probabilistic appearance manifolds. In Computer Vision and Pattern Recognition. Proceedings. IEEE Computer Society Conference.

G. Aggarwal, A. K. R. Chowdhury and R. Chellappa, 2004. A system identification approach for video-based face recognition. In Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04). IEEE Computer Society.

O. Arandjelovi´c and R. Cipolla, 2009. A pose-wise linear illumination manifold model for face recognition using video. Computer vision and image understanding, 113(1).

M. Nishiyama, O. Yamaguchi and K. Fukui2005. Face recognition with the multiple constrained mutual subspace method. In audio-and video-based biometric person authentication. Springer.

J. Li, Y. Wang, and T. Tan, 2005. Video-based face recognition using a metric of average euclidean distance. Advances in biometric person authentication.

J. Suneetha, 2014. A survey on video-based face recognition approaches. International journal of application or innovation in engineering & management, 3(2), (IJAIEM).

Z. Zhang, Chao Wang and Yunhong Wang, 2011. Video-Based Face Recognition: State of the Art, Lecture Notes in Computer Science: Biometric Recognition, (CCBR).

L. Best-Rowden, B. Klare, J. Klontz, and A. Jain, 2013. Video-to-Video face matching: Establishing a baseline for unconstrained face recognition. Sixth Int. Conference on biom. Compe. IEEE, (BTAS).

S. A. Patil and Paramod j Deore, 2012. Video-based face recognition: a survey. Proceedings of Conference on Advances in Communication and Computing (NCACC'12).

K. Khurana and B. Chandak, 2013. Key frame extraction methodology for video annotation. International journal of computer engineering & Technology, 4(2), (IJCET).

G. Liu, and J. Zhao, 2009. Key frame extraction from MPEG video stream. Proceedings of the second symposium international computer science and computational technology (ISCSCT'09).

T. Jabid, M. H. Kabir, and O. S. Chae, 2010. Local directional pattern (LDP) for face recognition. Proc. IEEE Int. Conference of Consumer Electronics.

T. Jabid, M. H. Kabir, and O. S. Chae, 2010. Robust facial expression recognition based on local directional pattern. ETRI Journal 32(5).

D.J. Kim, S.H. Lee, and M.K. Sohn, 2013. Face recognition via local directional pattern. International Journal of Security and Its Applications. Papers 7(2).

M. Yang, P. Zhu, L. V. Gool, L. Zhang, 2013. Face recognition based on regularized nearest points between image sets. In IEEE FG.

M. Kim, S. Kumar, V. Pavlovic and H. Rowley, 2008. Face tracking and recognition with visual constraints in real-world videos. In Proc. CVPR.

P. Viola and M. J. Jones, 2004. Robust real-time face detection. Int. journal of computer vision, 57(2).