Frontal Facial Pose Recognition using a Discriminant Splitting Feature Extraction Procedure

I. Marras[1,2], N. Nikolaidis[1,2], and I. Pitas[1,2]
[1]*Department of Informatics, Aristotle University of Thessaloniki, Greece*
[2]*Informatics and Telematics Institute, CERTH, Greece*
E-mail(s): { imarras, nikolaid, pitas}@aiia.csd.auth.gr

Abstract. *Frontal facial pose recognition deals with classifying facial images into two-classes: frontal and non-frontal. Recognition of frontal poses is required as a preprocessing step to face analysis algorithms (e.g. face or facial expression recognition) that can operate only on frontal views. A novel frontal facial pose recognition technique that is based on discriminant image splitting for feature extraction is presented in this paper. Spatially homogeneous and discriminant regions for each facial class are produced. The classical image splitting technique is used in order to determine those regions. Thus, each facial class is characterized by a unique region pattern which consist of homogeneous and discriminant 2-D regions. The mean intensities of these regions are used as features for the classification task. The proposed method has been tested on data from the XM2VTS facial database with very satisfactory results.*

Keywords. Frontal Facial Pose Recognition, Facial Image Analysis, Semantic Video Analysis, Discriminant Image Splitting, Pose Estimation.

1. Introduction

Facial image analysis tasks such as face detection, clustering and tracking, facial features detection, face recognition or verification and facial expression recognition have attracted the interest of computer vision and pattern recognition communities over the past years due to their importance in a number of applications that include semantic video analysis and annotation for archival, indexing and content management, 3D reconstruction, human-computer interaction etc. In some of these facial tasks such as face recognition and facial expressions recognition, the majority of developed techniques have been designed to operate on frontal or nearly frontal face images [12,1,11]. Due to this fact, a face or facial expression classifier trained on frontal facial images will not be able to operate successfully on non-frontal images. As a result, techniques that recognize frontal facial poses need to be developed, so that frontal facial images can be selected among all available facial images and used as input in face recognition or facial expression recognition systems. The same problem arises in cases where multiple view video data, acquired through a convergent multi-camera setup, are available. In this case, a frontal facial pose recognition algorithm can be applied on the available video streams to identify the view that is closer to a frontal one. By doing so, frontal images of the person can be acquired and fed to a face or facial expression recognition technique that requires frontal faces, leading to view-independent recognition.

Obviously, one can find frontal facial images by using head pose estimation techniques [6], that estimate the orientation of the head by determining the value of the yaw, roll and pitch angles. However, this would introduce unnecessary complexity and load to the overall face/facial expression recognition system since in this case one only needs to distinguish frontal over non-frontal poses and not obtain a detailed head orientation estimate. The method proposed in [3] uses a novel pose estimation algorithm based on mutual information to extract any required facial poses from video sequences. The method extracts the poses automatically and classifies them according to view angle. The method proposed in [4] firstly employs the Discriminant Non-Negative Matrix Factorization (DNMF) algorithm on the input images acquired

from every camera. The output of the algorithm is then used as an input to a Support Vector Machines (SVMs) system that classifies the head poses acquired from the cameras to two classes that correspond to the frontal or non frontal pose. In this paper, we view the frontal face pose recognition problem in a simplified framework. Rather than determining the head pose angles, we simply consider an image of a face to be either frontal or non-frontal. This is essentially a two-class problem (frontal vs non-frontal). However, the fact that the non-frontal class is much richer, since it contains all possible head orientations except for the frontal one, led us to split the non-frontal class into a number of classes, each containing non-frontal images where the head orientation lies within a range of angle values. Obviously all facial images classified to one of the non-frontal classes are labelled as non-frontal.

Motivated by the use of graphs with nodes placed at discriminant points for face recognition [10], our frontal facial pose recognition technique segments the facial images to discriminant regions. The main idea is the creation of a set of regions that is discriminative for each class of facial images in the sense that a subset of these discriminant and homogeneous regions will provide adequate information in order to distinguish a certain class from another one. The entire set is necessary in order to distinguish this class from the rest of the other classes. The region segmentation is based on the classical image splitting technique. The features that this method uses are the mean intensities of the produced regions. In general, the proposed method constitutes the basis for biometric techniques developed in the future.

This paper is organized as follows: the discriminant splitting approach used in the training phase in order to extract the characteristic features for each facial class is presented in Section 2. The actual frontal pose recognition (classification) procedure and the experimental evaluation of the method on the XM2VTS facial database are presented in Section 3 and 4. Section 5 provides conclusions.

2. Feature Extraction Using Discriminant Splitting

Let us assume that there exist *n* facial image classes namely one class containing frontal facial views (including small deviations from the fully frontal view) and *n*-1 classes corresponding to non-frontal views. Each class contains *l* different, equally sized, training aligned facial images. Each of the non-frontal classes contains images where the head rotation angle with respect to the vertical axis (yaw) lies within a certain interval. Thus, the dataset *D* is divided into *n* sets $D = \bigcup_{i=1}^{n} \boldsymbol{u}_i$. The main goal is to find homogeneous regions that are discriminant between the classes. In this way, for each class, a unique regions pattern, i.e. a set of regions, is created. This procedure, that is based on a splitting approach, will be described bellow.

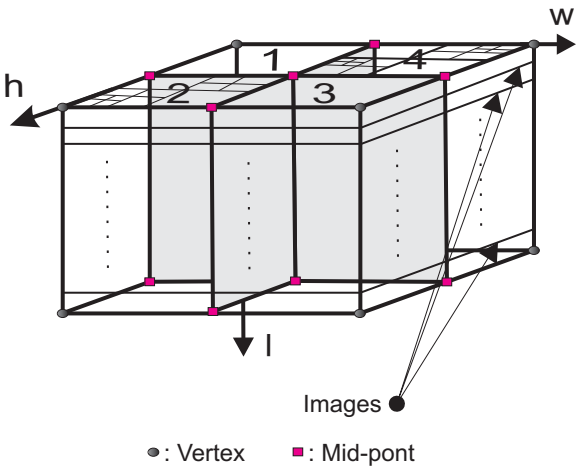

Figure 1. Class representation as a stack of images.

Let two classes *a*, *b* each containing *l* samples (images) of the corresponding facial class, in sets $\boldsymbol{u}_a$, $\boldsymbol{u}_b$. If each image is of dimensions $h \times w$, these *l* images can be considered as a stack of slices (volume) with dimensions $l \times h \times w$. Thus for our purpose, a certain region *B* can be considered as being a parallelepiped volume comprising of the parts of every image in the class that fall within the region, as illustrated in Fig. 1. We assume that an image *I* is divided into *R* regions. For a region *B* defined as above and for a class *a* we define its discriminant power, with respect to class *b*, using the Fisher's discriminant ratio [2] that is:

$$F_{a,b}(B) = \frac{(\mu_a(B) - \mu_b(B))^2}{\sigma_a(B)^2 + \sigma_b(B)^2}, \qquad (1)$$

where $\mu_a(B)$ and $\sigma_a(B)$ are the mean intensity and variance for the region *B* of all the images that belong to class *a*, while $\mu_b(B)$ and $\sigma_b(B)$

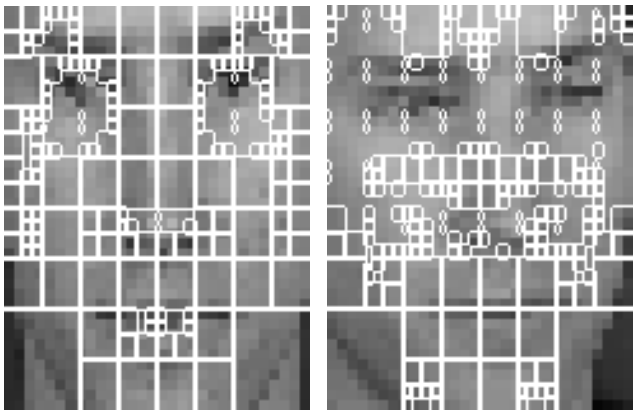

Figure 2. Facial images from the XM2VTS database that belong to the frontal class and one of the non-frontal classes, along with the corresponding region patterns.

are the mean intensity and variance for the same region of the images that belong to class *b*. A region B_1 is more discriminant than a region B_2, for a particular pair *a*, *b* of facial classes, when $F_{a,b}(B_1) > F_{a,b}(B_2)$. As mentioned above, except from the discriminant power of a region the method exploits also its homogeneity. Any volume homogeneity measure can be used. We have chosen the one based on the intensity range $|I_{\max} - I_{\min}|$, where $I_{\max}$, $I_{\min}$ are the maximum and minimum intensity values of a region. If the range is smaller than a certain threshold, i.e.:

$$|I_{max} - I_{min}| \leq T_s \qquad (2)$$

then the region is regarded to be homogeneous, where the threshold T_s denotes the Otsu threshold [7] calculated for the current region. As in the case of region discriminant power calculation, the homogeneity of a region is judged based on the pixels intensity values of the parts of *all* the class's training images that fall within the region's boundaries, i.e. on all pixels of the corresponding volume.

In order for the discriminant and homogeneous regions to be determined for each class *a*, the classical splitting approach is applied to the *l* images of this class. The corresponding stack of images is recursively split into four quadrants or regions (Fig. 1), until 2D discriminant and non-homogeneous regions are encountered. The splitting is performed by bisecting the rectangular regions (in the entire image stack) in the vertical and horizontal directions.

The splitting procedure for the stack of slices of a class *a* proceeds as follows: For a region *B* under consideration we evaluate the discriminant ratio in (1) for all pairs of classes, i.e. we evaluate $F_{i,j}(B)$, $i, j = 1 \ldots n$, $i \neq j$. We then find the largest ratio $F_{\max}(B) = F_{i^*,j^*}(B) = \max_{i,j}(F_{i,j}(B))$. If this maximum involves the region *B* in class *a*, i.e. if $i^* = a$ or $j^* = a$ this means that *B* is most discriminant in the task of distinguishing *a* from the other classes. In this case, region *B* is split for class *a*. Otherwise, the homogeneity of *B* in *a* is examined using (2) and the region is split if it is inhomogeneous, until homogeneous regions are reached, otherwise it is not split. In summary, if a region is very discriminant for a class it is being split, whereas if it is not discriminant enough, it is split if it is inhomogeneous.

The above procedure is performed for each facial class separately. In the end of the training procedure, a region pattern for each class is created. Two facial images that belong to different classes along with the corresponding region patterns are shown in Fig. 2.

Finally, each training image I_k within a class a is characterized by the vector $\underline{\boldsymbol{\mu}}_{ak}$ that contains the mean intensity values for each of the regions $r_{a,j}$, $j = 1 \ldots n_a$, n_a being the number of regions in the pattern of class *a*.

Figure 3. A frame from the XM2VTS database.

Figure 4. Frontal (top row) and non-frontal (bottom row) facial images from the XM2VTS database.

3. Image Classification

The algorithm's testing (classification) procedure is as follows. The n_a discriminant regions r_{aj} of class *a* are selected upon an image *I* depicting a face in an unknown orientation. In order to solve small alignment and scaling problems, the pattern (set) of regions for each class is considered as an elastic (non-rigid) pattern, so, the regions boundaries are translated locally by small amounts until they fall on, as much as possible, homogeneous regions. For a class *a*, the intensity means $\boldsymbol{\mu}_{Ir_{aj}}$ of every region r_{aj} of class *a* are computed in *I*, providing the image *I* means vector $\underline{\boldsymbol{\mu}}_{Ia}$. The image means vector $\underline{\boldsymbol{\mu}}_{Ia}$ is then compared with the (pre-computed) means vectors $\underline{\boldsymbol{\mu}}_{ak}$ for all training images I_k ($k = 1 \ldots l$) of facial class *a*, resulting in distances $d_{Ia_k} = \left\| \underline{\boldsymbol{\mu}}_{Ia} - \underline{\boldsymbol{\mu}}_{ak} \right\|$ for every training image *k* that belongs to class *a*. Thus, *l* distances are computed for each class. This is repeated for all n_{Total} classes resulting into $l * n_{Total}$ distances. The facial image is classified to the class a^*, that contains the training image k^* whose means vector is closest to the test image means vector,

$$(a^*, k^*) = \arg\min_{a,k} d_{Ia_k}. \qquad (3)$$

The small region translations mentioned above are also involved in this search for the matching training image and class.

In literature, many methods have been proposed that are robust to light variations [8]. The proposed method is robust to small light variations as those appearing in XM2VTS face database, while in other cases, techniques for removing illumination artifacts can be used [9], or an extra sub-class could be created for each different light direction.

4. Experimental results

The proposed method was evaluated on data obtained from the XM2VTS face database [5]. For testing the capability of the algorithm having many persons and different video sessions of them, we used the XM2VTS video database. This database contains four recordings of 295 subjects taken over a period of four months. Each recording is comprised of a speaking head shot and a rotating head shot. Sets of data taken from this database are available including high-quality color images, 32 kHz 16-bit sound files, video sequences, and a 3D model. In the first shot of each session, the person reads a given text, while in the second shot each person moves its head in all possible directions allowing multiple views (poses) of the face.

Face tracking was applied on the head rotation shot videos, depicting people that start from a frontal pose, turn their heads to their right extreme, back to frontal pose then to the left extreme (Fig. 3). The resulting facial images, that depict the face bounding box (Fig. 4), were then rescaled to a size of 30×40. 6862 facial images were obtained, 2486 being frontal and 4376 non-frontal. Images where the head rotation is in the range $[-10^{\circ}\ldots10^{\circ}]$, zero degrees being the frontal orientation, were considered as frontal. The non-frontal images were split into four classes containing facial images with head orientations in the ranges $[10^{\circ}\ldots50^{\circ}]$, $[50^{\circ}\ldots90^{\circ}]$, $[-10^{\circ}\ldots-50^{\circ}]$, $[-50^{\circ}\ldots-90^{\circ}]$. Each of these four classes contained roughly the same number of images. We then randomly split the images in half for all five classes to form the training and test sets. Thus, both sets consisted of 1243 frontal face images and 2188 non-frontal images with no overlaps between the sets.

The proposed algorithm was found to be able to classify facial images to frontal and non-frontal with very satisfactory accuracy. Indeed the correct classification percentage achieved by the proposed method was 98.8%, while the method reported in [3] achieves correct classification percentage 98.1%. The mis-classifications were because we have discretized a continuous regression problem. Table 1 shows the confusion matrices for the method proposed in [3] and the proposed method, in XM2VTS face database. In terms of computational complexity, 2.2 seconds are required for the proposed algorithm in order to categorize a facial image on an Intel Pentium 4 (3.01 GHz) processor PC with 1.5GB of RAM.

Table 1. The confusion matrices for the method proposed in [3] and the proposed method, in XM2VTS face database.

Method proposed in [3]		
	Frontal	**Non-Frontal**
Frontal	**2455**	**99**
Non-Frontal	**31**	**4277**

Proposed method		
	Frontal	**Non-Frontal**
Frontal	**2469**	**65**
Non-Frontal	**17**	**4311**

5. Conclusion

A novel method for frontal facial pose recognition that is based on discriminant region splitting is proposed in this paper. Non-overlapping regions form a unique region pattern for each class, namely the frontal class and the classes that correspond to non-frontal views. The mean intensity values of each of the regions in this pattern are used to characterize this class and classify an image depicting an unknown facial orientation. Results show that the proposed technique is able to classify facial images to frontal and non-frontal with very good accuracy. It should be noted that the proposed feature extraction and classification procedures can be also used in similar classification tasks.

6. Acknowledgements

The research leading to these results has received funding from the European Community Seventh Framework Programme (FP7/2007-

2013) under grant agreement no 211471 (i3DPost).

7. References

[1] Fasel B, Luettin J. Automatic facial expression analysis: A survey. *Pattern Recognition*, 36:259-275, 1999.

[2] Fukunaga K. Statistical Pattern Recognition. *Academic Press*, San Diego, CA, 1990.

[3] Goudelis G, Tefas A, Pitas I. Automated facial pose extraction from video sequences based on mutual information. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(3):418-424, March, 2008.

[4] Kotsia I, Nikolaidis N, PitasI. Frontal view recognition in multiview video sequences. *In Proceedings of IEEE International Conference on Multimedia and Expo (ICME'09)*, pages 702-705, 2009.

[5] Messer K, Matas J, Kittler J, Jonsson K. XM2VTSDB: The extended M2VTS database. *In 2nd International Conference on Audio and Video-based Biometric Person Authentication*, pages 72-77, 1999.

[6] Murphy-Chutorian E, Trivedi M. M. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607-626, April 2008.

[7] Otsu N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9:62-66, 1979.

[8] Tzimiropoulos G, Zafeiriou S, Pantic M. Principal component analysis of image gradient orientations for face recognition. *In Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, Santa Barbara, CA, USA, March 2011.

[9] Struc V, Vesnicer B, Mihelic F, Pavesic N. Removing illumination artifacts from face images using the nuisance attribute projection. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'10)*, pages 846-849, Dallas, Texas, USA, March 2010.

[10] Zafeiriou S, Tefas A, Pitas I. Learning discriminant person-specific facial models using expandable graphs. *IEEE Transactions on Information Forensics and Security*, 2(1):55-68, 2007.

[11] Zeng Z, Pantic M, Roisman G. I. Huang T.S. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39-58, January 2009.

[12] Zhao W, Chellappa R, Rosenfeld A, Phillips P. J. Face recognition: A literature survey. *ACM Computing Surveys*, pages 399-458, 2003.