Hand detection in cluttered scene images using Fourier-Mellin invariant features

Lluís Gómez i Bigordà, David Masip, Universitat Oberta de Catalunya

Abstract — This paper proposes an automatic hand detection system that combines the Fourier-Mellin Transform along with other computer vision techniques to achieve hand detection in cluttered scene color images. The proposed system uses the Fourier-Mellin Transform as an invariant feature extractor to perform RST invariant hand detection. In a first stage of the system a simple non-adaptive skin color-based image segmentation and an interest point detector based on corners are used in order to identify regions of interest that contains possible matches. A sliding window algorithm is then used to scan the image at different scales performing the FMT calculations only in the previously detected regions of interest and comparing the extracted FM descriptor of the windows with a hand descriptors database obtained from a train image set. The results of the performed experiments suggest the use of Fourier-Mellin invariant features as a promising approach for automatic hand detection.

Index Terms — Automatic hand detection, Fourier-Mellin Transform, RST-invariant object representation.

- *Ll. Gómez i Bigordà is a Master in Computer Science student in the Universitat Oberta de Catalunya, Barcelona, Spain. E-mail: lgomezbi@cv.uoc.edu.*
- *D. Masip is with the Department of Computer Science, Multimedia, and Telecommunications, Universitat Oberta de Catalunya, Rambla del Poblenou 156, 08018 Barcelona, Spain. E-mail: dmasipr@uoc.edu.*

1 Introduction

AUTOMATIC recognition of objects in uncontrolled environments remains an important open problem in the area of artificial intelligence. There are countless areas of research with great interest in the techniques of object recognition, for example in the fields of image indexation, digital watermarking, face detection and facial recognition, character recognition, tracking and counting objects, stabilizing video signals, image resgistration and much more. All these cases are an evidence that object recognition have a central role in present and future applications of computer vision in our daily lives.

The state of the art on object recognition shows a growing tendency to use methods that combine techniques from different disciplines, such as artificial intelligence and machine learning, statistical analysis, signal processing or image filtering, resulting in a number of different hybrid methods that address the problem of object recognition with varying efficiency depending on how the problem to solve is defined. Thus, for example, there are methods and procedures that work best in face detection and others do for character recognition. There is no single method that can solve the problem of object recognition in general terms in a 100% effective way.

Hand detection, and other objects which can exist in a infinite bunch of different forms in an image, is one of the most difficult recognition tasks to achieve. One of the main problems of automatic recognition of objects in general, and specifically on hand detection, is the great diversity of different versions in which the same object can be presented. These differences between objects of the same class can not only come from the inherit inter-class variability, nor to changes in the characteristics of the image, such as lighting conditions or the viewpoint of the observer, but also from the fact that we can find the object in any position within the image, or at different scales and rotations. This highly complicates the recognition of objects because it is very difficult to establish a correct correspondence between pixels in different images.

The algorithms used for object recognition mainly differ in the features on which the recognition is based. There are a lot of effective methods for extracting features of an image, methods in the state of the art include the Histograms of Oriented Gradients (HOG) [1], the Haar-like features, the Fourier-Mellin Transform [2] – [5], which is sometimes modified in order to achieve a better performance [6] - [8], along with a long list of invariant moments such as Hu, Legendre, Zernike, Pseudo-Zernike, Chebyshev, or the own Fourier-Mellin moments [2], [9].

The present research focus on the Fourier-Mellin Transform as a feature extraction method. The method used here for calculating the Fourier-Mellin Transform [4] has been used successfully on numerous occasions. In the state of the art of object recognition are many and diverse applications of the Fourier-Mellin Transform based descriptors, from "synthetic" environments which simply seek to demonstrate the invariance properties of the Fourier-Mellin Transform [7] [8] to very efficient methods in face recognition with one example image per person [4], to search handwritten documents [6], detection of digital watermarks, registration (or alignment) of satellite images, image spam filtering [5] or face detection [9].

As the aim of this paper is mainly to search for evidence of the feasibility of hand detection using Fourier-Mellin Transform based descriptors in the simplest

method possible, sophisticated supervised machine learning classification algorithms are not used here. Although these shall be logical in this kind of object detection tasks, We understand those are more appropriate for a future work. As an alternative to state-of-the-art machine learning classifiers We use a brute force search algorithm based on sliding window, which despite being computationally expensive offers a simple way to check the validity of the Fourier-Mellin feature descriptors, and at the same time allows us to avoid the complexity and long lasting process of sophisticated classifiers training. Our approach consists in a first stage where a set of training sample images are analyzed to create a database of Fourier-Mellin descriptors, and a second stage where the sliding window algorithm search for matches comparing the FM descriptor of all possible windows in a test image with the patterns stored in the database.

Two other computer vision techniques of easy implementation have been used in combination with the sliding window algorithm in order to reduce computational cost: a simple skin color-based segmentation and a well known interest points detector, which have helped improving both the computing cost and the efficiency of the method.

2 Implementation

2.1 Fourier-Mellin invariant feature extractor

In general terms, the basic idea behind the Fourier-Mellin Transform is to convert the changes in rotation, scale and translation (RST) to translations on any axis of the transformed image, as that translations can be properly "corrected" using the magnitude of the Fourier Transform.

It is well known that the rotation of an image at an angle ϕ causes a rotation in the Fourier Transform of the same angle and in the same direction, and that changing spatial scale in an image by a factor ρ cause a change of scale in the magnitude of its Fourier Transform by a factor $1/\rho^2$ and in frequency by a factor $1/\rho$ [7].

An invariant descriptor based on the Fourier-Mellin Transform can be obtained with the following chain of transforms [4], [6], [7]:

- First estimate the magnitude of the 2D Discrete Fourier Transform (DFT) of the original image.

$$|F\{I[m,n]\}| = |F[k,l]| = \left| \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I[m,n] e^{-j(2\pi/M)(km)} e^{-j(2\pi/N)(\ln)} \right| \quad (1)$$

- Then apply a mapping from Cartesian coordinates to polar logarithmic coordinates.

$$|G\{F[x,y]\}| = G[\ln(r), \theta] \quad (2)$$

$$r = \sqrt{x^2 + y^2}$$

$$\theta = \arctan\left(\frac{y}{x}\right)$$

- Finally, estimate the magnitude of the Discrete Fourier Transform (1) of the matrix resulting from the previous step.

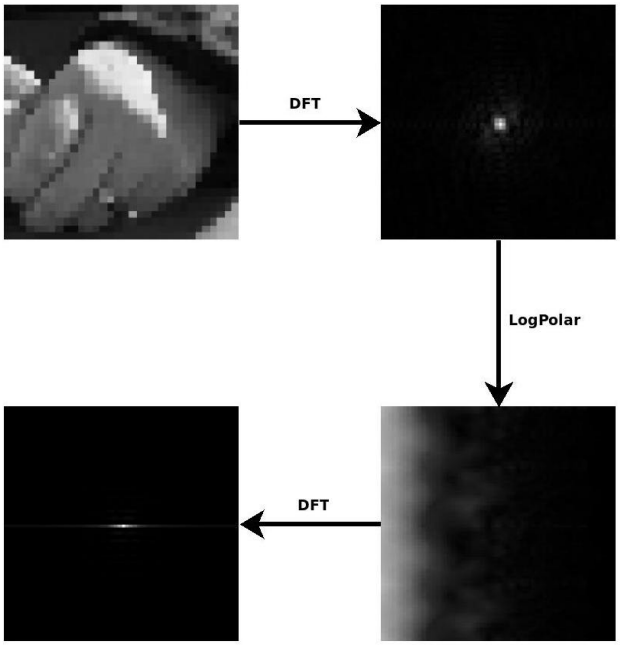

Figure 1: Stages to obtain an invariant descriptor based on the Fourier-Mellin Transform

The Fourier-Mellin Transform itself is not scale invariant, but one can add some scale invariance with a simple normalization procedure on the outcome of log-polar coordinates mapping previously described [7] resulting in a useful modification of the Fourier-Mellin Transform[6]. However, this will not be the case in the present method as in our case the image will be scanned using a sliding window algorithm where all possible window sizes will be analyzed. Thus implies to have all the FM descriptors in our database in a fixed resolution, and then convert all possible windows scanned to same scale before to do the FM Transform.

Obviously, the performance of the descriptors obtained that way may vary depending on the resolution on which we decide to work. Finest resolutions for the first DFT achieves better performance but at the same time increases computational cost exponentially. In order to obtain an acceptable balance between precision and computational cost a first set of experiments have been done (see Experiments section below). The experiments concluded that taking the first DFT (1) at a 64x64 resolution achieve a considerable increase of performance that can't be underestimated in order to obtain an acceptable detection rate in the final detector.

However, a 64x64 size for a feature descriptor is an undesirable big size for our purposes as, although the FM Transform can be fast computed for that size, We'll be performing lots of FM descriptor comparisons when we scan

an image with the sliding window algorithm. In order to reduce the dimensions of our feature descriptors We can take advantage of an important property of the Fourier Transform: the magnitude of low frequencies contain more image information than the higher ones. Several low-pass filters has been tested in order to reduce the dimension of our descriptors without loosing information, thus We added a new step in our feature extractor which crops the last obtained DFT magnitude to a 16x16 matrix in the center. This dimensionality reduction of the descriptors notably reduces computational cost in the sliding window procedure achieving same detection rate as with the unfiltered descriptors.

Finally, a last improvement is introduced to our feature extractor to remove differences introduced by rotating the same image over a fixed analysis window:

Figure 2: A squared window of the original image (on the left) and on the rotated version (right) doesn't contain the same information.

As shown in the above images the rotated image shows informations in the corners of the window that where not present in the original image, at the time that hides the pixels originally in those corners. Applying a circular mask to the input image of our feature extractor those differences disappear, and thus the FM feature descriptors of both images will match better:

Figure 3: Applying a circular mask to the windows on both original image (left) and the rotated version (right) the information shown is now practically identical.

Although the mask improvement fits better for a 2-D object detection system, has been proved to be useful in our experiments where the training descriptors comes from an image set where hands are annotated as bounding boxes that likely introduce distracting information on the corners.

The stages of our final Fourier-Mellin feature extractor are shown in the image below:

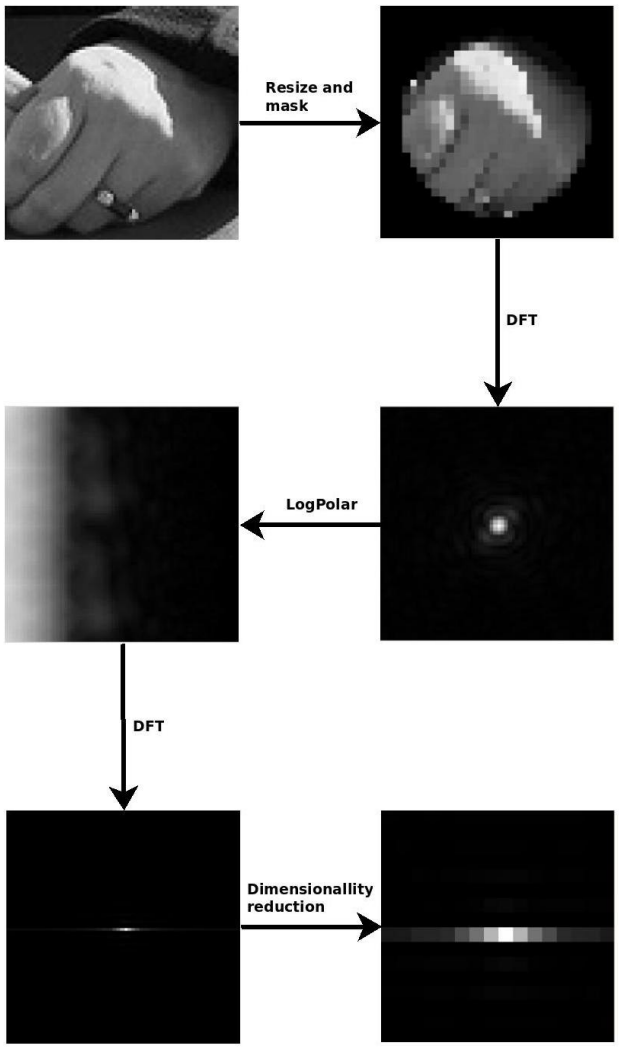

Figure 4: Stages of our Fourier-Mellin feature extractor.

2.2 Sliding window search method

Sliding window algorithms have been a commonly used method for many years in the field of object detection. Those are simply based on evaluation of a score classifier in all possible rectangular subregions of an image, the maximal scores found indicates possible locations for the searched object. As the the number of unique rectangles in an N×N image is of the order N^4, the exhaustive task is in general computationally temporarily intractable, and thus impossible to be done.

In our particular case, the score classifier computes the distance between the local FMT descriptor and each hand FMT descriptors in our database:

$$distance = \sqrt{\sum_{x,y=0}^{x,y=N} (i_{x,y} - j_{x,y})^2} \quad (3)$$

So, the exhaustive search task implies N^4 MFT feature extractions and comparing of all them with a 300 descriptors database.

In order to speed up the search process, several heuris-

tics which reduce the number of evaluated windows have been applied:

- Search only inside human bounding boxes provided for the image test set.
- Allow only perfect square windows of certain fixed sizes.

On the other hand, local optimization methods have been used to identify regions of interest in the image where a hand is more probable to be found: a simple skin color-based segmentation and a well known interest points detector. Although this techniques reduced considerably the computation cost of our sliding window algorithm it's also true that can lead to false detections or even complete misses of the searched hands, for example when skin color is not well classified.

2.3 Skin color-based segmentation

The simplest skin color classifiers are based on delimiting the subspace of skin color within the color space in which them work, for example in the RGB color space the simplest skin color-based classifier provides an upper and lower threshold for each color component (Red, Green and Blue) so that if a pixel has all the RGB values between this thresholds is considered to be skin.

To obtain certain resistance to different lighting conditions, such as the color of the light and shadows, the chromatic color space or "normalized RGB" [10], [11] is used. Chromatic colors also known as "pure colors in absence of luminance", are defined by a normalization process that uses a proportional part of each color:

$$base = R + G + B$$

$$R_c = R / base \qquad (4)$$
$$G_c = G / base$$

(4) Definition of the chromatic color space.

Note that Bc is omitted because its value is redundant:

$$R_c + G_c + B_c = 1 \qquad (5)$$

The chromatic color space provides a commonly known and simple way to quickly detect skin color pixels via the following classifier:

$$Skin = (0.5 > R_c > 0.35) \wedge (0.7 > G_c > 0.2) \wedge (base > 200) \qquad (6)$$

(6) Simple skin chromatic color-based classifier.

Using the simple skin chromatic color-based classifier (6) as a boolean pixel operation a skin segmentation is performed in order to identify the regions of interest where the sliding window algorithm will perform FMT calculations and comparisons.

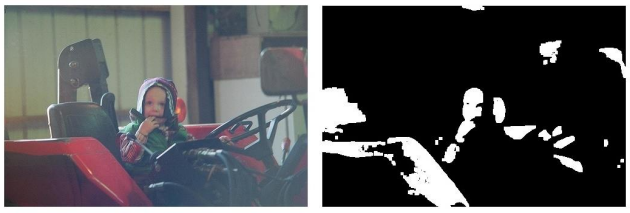

Figure 5: An image from the train set (left) and the output image (right) of the skin segmentation operator (6).

2.4 Interest points detector

In general terms the concept of interest points relies on the idea that some points exist in an image that are distinguishing and stable features that can be accurately localized. The classical approach to detect interest points of an image is based on the Harris corner algorithm[12]. In our system the Shi & Tomasi[13] variation is used, as it's implemented in several programming languages and platforms and widely available.

For their own nature, hands usually presents sharp shapes that often match with interest points of the image. We have found that around 90% of the 351 hands in our train image set have at least one interest point inside their bounding boxes.

Thus, the interest points detector is the second optimization method used as a regions of interest finder for the sliding window algorithm, but also is used to purge the FMT feature descriptors database removing the hands that don't match with interest points, considering them as bad hand "examples" for our purposes.

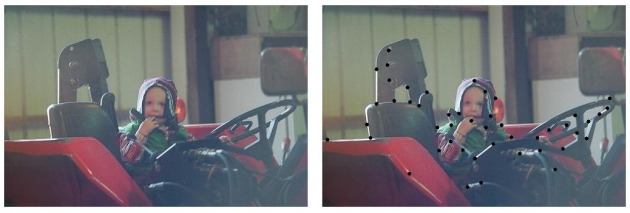

Figure 6: An image from the train set (left) and the output image (right) of the interest points detector.

3 Experiments

3.1 Fourier-Mellin invariant feature analysis

A first experiment was designed in order to verify the invariant properties of our FMT feature extractor. Only images from the train set where used in this experiment. One by one a collection of train images where modified using an automated tool to randomly perform RST variations, the modified image was then fully scanned with the sliding window algorithm. Rotations where performed in any angle from 0 to 360 degrees, and scaling at different form factors from 0.8 to 2. Initially smaller form factors where allowed but the method was unable to work properly with very small windows (p.e. 10x10) resulting for example from the 0.5 downsizing of an image with a hand

in a 20x20 bounding box, which is a frequent size in our data set. However, this limitation is not a weakness of the FMT, but rather a threshold for the minimum information necessary to recognize an object, because correct detections were obtained downsizing with a scale factor of 0.5 in images with larger hands.

We tested the method in a set of 20 images, each one randomly transformed in 5 different ways, so the total amount of scanned images was 100.

When the sliding window is performed with only one descriptor in the database (the one obtained from the scanned image in its original version), the achieved hand detection was 98%. It's important to appreciate that (by chance) none of the images in this set have any hand in regions classified as non-skin by the skin color-based classifier, otherwise, of course, the number of misses would have increased.

Several rotated images with correct hand detection are shown below (green boxes means the minimal value of the distance (3), red ones are other low-distance windows):

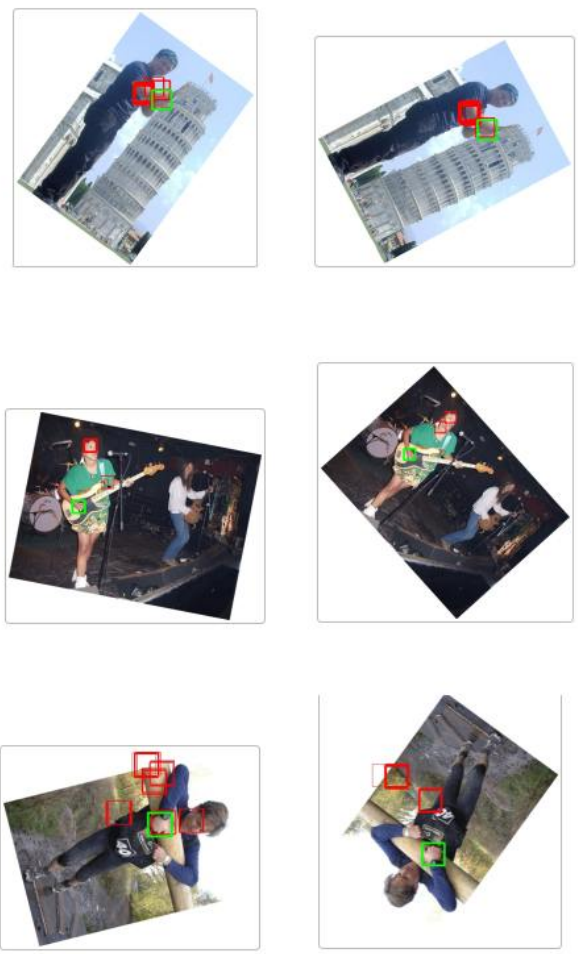

Figure 7: Correct hand detections on rotated images from the train set.

Things change to worse when the full database is used, so when all the 307 FMT descriptors extracted from the training set are compared one by one with each possible windows. In this case the hand detection rate decrease around 22%. Most of the misleading detections involve distraction zones where, despite not having a strong hand appearance, the FMT distance score classifier concentrates maximal values in a surprising manner. Next figure shows an evidence of this behavior:

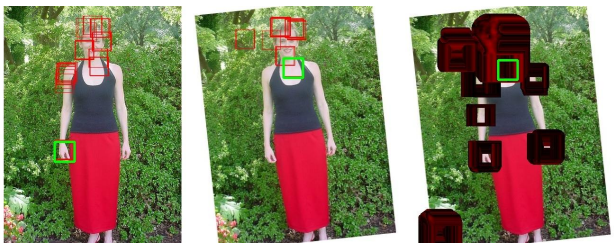

Figure 8: Correct hand detection on an image from the test set(left), misleading detection on a rotated version(center), and hand detection probabilistic representation (right) of the rotated image.

The image on the left of Figure 8 is the original image as it is in the training set, and obviously the detector works perfect there (with distance (3) equals to zero), the second image is a misleading classification for a 354 degrees rotated variation, the classifier not only gets the minimal distance (3) far from the hand but also the hand is totally ignored for all the other possible matches shown in red. The third image in Figure 8 is a probabilistic representation of the score function of our classifier: alls windows with a distance under a quite permissive threshold are drawn in red tones, the darkness ones are the less probable matches and the lighter red ones are the closest to some FM descriptor in the database (i.e. distance (3) is closest to zero). The big concentration of lighter red boxes around the head of the woman character somehow suggest the existence of weak descriptors in the database that should be purged in some way in order to achieve better results. This, however, is out of the scope for this experiment, but remains as one of the possible extensions of the presented method.

3.2 A real world hand detector

Being demonstrated a sufficient robustness for our FM invariant feature extractor another experiment was performed in order to test the method in a real world hand detection task. In this case the images of the train set are only used to feed the FM descriptors database which the sliding window algorithm uses to calculate distances (3) in the score classifier.

No previous information is revealed about the input images for this experiment, any kind of color images independently on how cluttered their scenes are, with absolutely no relation with the train images, can be part of the test set. Thus this is a real world application and, as stated before, hand detection in this conditions is one of the most complicated tasks in the field of object recognition.

Images with misleading skin detection or without interest points on the hands present where removed from the test set and are not part of the final detection rate. The

proposed method was applied into an alternate set of 86 images achieving a hand detection rate of 9'3%. Again the misleading detections seems to be introduced by weak descriptors in the database.

Figure 9: Some of the hands detected with the presented method.

Conclusion

Far from being a final solution, the proposed method shows sufficient evidence to be considered and explored in greater depth.
The robustness of the invariant properties of the FMT feature descriptor have been confirmed by the experiments and the rate of correct detections of the real world hand detector, although low, is not far from results obtained with other methods in the last PASCAL Challenge Workshop celebrated jointly with ECCCV 2010.

The proposed method can be improved in many directions, including but not limited to:

- Purge of weak descriptors, and construction of a database of characteristic hand FMT descriptors.
- Use of a FMT descriptors pyramid at different resolutions to achieve more accurate detection.
- Improve the segmentation of the skin with an adaptive skin-color classifier.
- Use of state of the art supervised machine learning classification algorithms.
- Reduction of the computing cost.

Acknowledgment

The authors wish to thank Universitat Oberta de Catalunya for the resources and facilities offered. This work was supported in part by a grant from MEC.

References

[1] Pedro F. Felzenszwalb, Ross B. Girshick, McAllester D., Ramanan D., "Object Detection with Discriminatively Trained Part-Based Models", IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1627-1645, 2010.

[2] Chen Y. M., Chiang J.-H., "Face recognition using combined multiple feature extraction based on Fourier-Mellin approach for single example image per person", Pattern Recognition Letters, Vol. 31, Issue 13, pp. 1833-1841, 2010.

[3] Gotze N., Driie S., Hartmann G., "Invariant Object Recognition with Discriminant Features based on Local Fast-FourierMellin Transform", IEEE International Conference on Neural Networks, pp. 948-951 vol.1, 2000.

[4] Van Den Dool R., "Image Processing Tools, Fourier-Mellin Transform", Stellenbosch University, (2011/06/19). Available: http://students.ee.sun.ac.za/~riaanvdd/Tools%20-%20Fourier-Mellin%20Transform.doc

[5] H.Q. Zuo, X. Li, O. Wu, W.M. Hu, G. Luo., "Image spam filtering using Fourier-Mellin invariant features", IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 849-852, 2009.

[6] Kountchev R., Todorov V., Kountcheva R., "Invariant Object Representation with Modified Mellin-Fourier Transform", 14th WSEAS International Conference on Computers, pp. 232-236 Volume I, 2010.

[7] Raman S. P., Desai U. B., "2-D Object Recognition using Fourier Mellin Transform and a MLP Network", IEEE International Conference on Neural Networks, pp. 2154-2156 vol.4, 1995.

[8] Wang X., Xiao B., Ma J.-F., Bi X.-L., "Scaling and rotation invariant analysis approach to object recognition based on Radon and Fourier–Mellin transforms", Pattern Recognition, Volume 40, Issue 12 pp. 3503-3508 , 2007.

[9] Terrillon J.-C., Shirazi M. N., McReynolds D., Sadek M., Sheng Y., Akamatsu S. , Yamamoto K., "Invariant Face Detection in Color Images Using Orthogonal Fourier-Mellin Moments and Support Vector Machines", Lecture Notes in Computer Science, Volume 2013/2001, pp. 83-92, 2001.

[10] Vezhnevets V., Sazonov V., Andreeva A., "A Survey on Pixel-Based Skin Color Detection Techniques", Graphics and Media Laboratory, Faculty of Computational Mathematics and Cybernetics, Moscow State University, Moscow, Russia. 2003.

[11] Wimmer M., Radig B., "Adaptive Skin Color Classificator", Int. Journal on Graphics, Vision and Image Processing, Special Issue on Biometrics, vol 2 pp. 39-42, 2006.

[12] Harris C., Stephens. M. J., "A combined corner and edge detector", by Alvey Vision Conference, pp. 147-152, 1988.

[13] Shi J., Tomasi C., "Good features to track", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 593-600, 1994.

Lluís Gómez I Bigordà is a Master in Computer Science student in the Universitat Oberta de Catalunya, where also achieved a Computer Science degree. He is currently working in Hangar, a visual arts research and production center in Barcelona.

David Masip works as a professor in the Universitat Oberta de Catalunya, in the Scene Understanding and Artificial Intelligence group from the Computer Science department. He holds a B.A. degree in computer science from the Autonomous University of Barcelona (UAB), a M.S. degree in computer vision in the Computer Vision Center (CVC), and a Ph.D. degree in computer science and artificial intelligence from the UAB. As a researcher he has authored more than 40 papers, and his main research is related to statistical pattern recognition, feature extraction methods, object detection, and object recognition, in particular face classification.