Facial Features Detection Robust to Pose, Illumination and Identity*

Nicolas Gourier Daniela Hall James L. Crowley
Prima Project, GRAVIR Laboratory, INRIA Rhône-Alpes, Montbonnot, FRANCE

Abstract

This paper addresses the problem of automatic detection of salient facial features. Face images are described using local normalized gaussian receptive fields. Face features are learned using a clustering of the Gaussian derivative responses. We have found that a single cluster provides a robust detector for salient facial features robust to pose, illumination and identity. In this paper we describe how this cluster is learned and which facial features have found to be salient.

1. Introduction

We are interested in automatically determining which facial features can be most reliably detected under variations in illumination, position, orientation and human identity. Our objective is to obtain a set of facial features that can serve as landmarks for tracking and recognition of facial expressions. We employ a fast, pixel level, detection algorithm to isolate and normalize the face region. Normalized face images are described by calculating a vector of scale-normalized Gaussian derivatives at each pixel. Salient facial features are detected using linear combinations of these descriptors. Such functions are learned using K-means clustering of the Gaussian derivative responses obtained from a set of training images. The resulting clusters specify linear combinations of Gaussian derivatives that act as detection functions for facial features that remain salient under variations in pose, illumination and identity.

2. Approaches to Facial Feature Detection

Facial feature detection may be performed using global or local features. A popular method for global analysis of face images is to project a normalized image into a linear subspace determined using a technique such as principal components analysis [11]. However, PCA is sensitive to head orientation. Alternatively, global analysis may be performed by measuring the relative position of anatomical facial structures such as the eyes and lips [1]. Nevertheless, global techniques tend to be sensitive to partial occlusions of facial features, and to the identity. Local approaches are less sensitive.

We define salient features as features that draw attention. Features isolated in a dense feature space are salient features [17]. Determining such local feature points can be performed by partitionning the face image into several regions, by using textons as in [7] or finding generic features [3, 9, 10].Facial features detection can be done using eigenfeatures [14], blobs [15] or saddle points and maxima of the luminance distribution [16]. But such descriptors are sensitive to illumination and provide too many points, which can lead to accumulation errors. Interest points are not robust to pose, and are not well adapted to deformable objects such as the human face.

Our objective is to design descriptors that are robust to illumination, scale and orientation. A way to obtain more generic points is to use local feature vectors. Gabor wavelets can be used to detect scale-invariant feature points, as presented in [2], [13], [6] and [8]. However, Gabor wavelets tend to be computionally expensive and have parameters that are difficult to adjust. Gaussian derivatives describe the appearance of neighborhoods and are an efficient means to compute scale and illumination robust local features. We adapt this method to detect facial salient features with interesting invariance properties.

Our approach is divided into several modules. First we employ a robust face tracker to detect and normalize the image of the face. As described in section 3, this step provides an important reduction in computation time. Further operations are performed on the normalized face image. We compute features for salient face regions with a learning process described in Section 4. Then we show and discuss our results in Section 5.

3. Face Image Normalization

We employ a robust video rate face tracker to focus processing on face regions. Our tracker uses pixel level detection of skin colored regions based on probability density function of chrominance [5].

*0-7803-8566-7/04/$20.00

3.1. Pixel Level Detection and Tracking using Skin Chrominance

To detect the face, we first detect skin regions in the image using intensity normalized color. The human face is a highly deformable surface and can be illuminated under several conditions. The color of image skin regions is determined by the product of the spectrum of skin pigments and illumination [19]. While such regions may have strong variations in intensity, chrominance will remain constant. We use an intensity invariant feature vector for each pixel by normalizing the Red and Green component and dividing by the intensity to obtain a vector (r,g). The ratio of the histogram of known skin regions divided by the histogram of the entire image provides a lookup table that converts a pixel of chrominance (r,g) into the probability $p(Pixel \in Skin|r,g)$ that the pixel is part of a skin region, as shown by Bayes rule. This lookup table gives us a direct relation (1) between intensity normalized color and probability.

$$p(Pix \in Skin|r,g) = \frac{p(r,g|Pix \in Skin)p(Pix \in Skin)}{p(r,g)} \tag{1}$$

The skin probability map is obtained by computing $p(Pixel \in Skin|r,g)$ for each pixel in a determined region. Face position and extent are tracked using a Kalman Filter. The first and second moments of the face are used to normalize the face position and orientation, as well as the size and resolution of the imagette that represents the face. The region of interest (ROI) for a face is maintained by a tracking process. In each image, the skin probability map is calculated within the region of interest predicted by using a zeroth order Kalman filter weighted by a Gaussian [5]. We estimate first and second moments with the following formulas (2):

$$\begin{aligned}
\mu_x &= \tfrac{1}{S}\textstyle\sum p_{skin}(x,y)\cdot x \cdot G(x,y,\vec{\mu},C),\\
\mu_y &= \tfrac{1}{S}\textstyle\sum p_{skin}(x,y)\cdot y \cdot G(x,y,\vec{\mu},C),\\
\sigma_x^2 &= \tfrac{1}{S}\textstyle\sum p_{skin}(x,y)(x-\mu_x)^2 G(x,y,\vec{\mu},C),\\
\sigma_y^2 &= \tfrac{1}{S}\textstyle\sum p_{skin}(x,y)(y-\mu_y)^2 G(x,y,\vec{\mu},C),\\
\sigma_{xy} &= \tfrac{1}{S}\textstyle\sum p_{skin}(x,y)(y-\mu_y)(x-\mu_x) G(x,y,\vec{\mu},C),
\end{aligned} \tag{2}$$

where $S = \sum p_{skin}(x,y)\cdot G(x,y,\vec{\mu},C)$

3.2. Performance of the Face Tracker

To initialize our face tracker, we employ either the user's selection on the frame, or a generic ratio histogram. The choice of the number of histogram cells is crucial to provide a confident skin probability map. Histograms with too few cells will not be discriminative enough to compute the skin probability, whereas histograms with too many cells can contain empty cells.

We achieve real time processing by avoiding image copy. The ROI is scanned only once. The face tracker runs at video-rate on Pentium 800 MHz with images of 384x288 pixels. Eye detection rate on representative video sequences can be seen in table 1 and Figure 1. In this case, an error occurs when the computed ellipse does not contain an eye visible in the image.

Table 1: *Eye detection rate*

Sequence	Number of images	Eye Detection rate
A	500	99,9 %
B	700	99,8 %
C	580	94,2 %
D	300	93,1 %

A : Head slow translation
B : Head fast translation
C : Head zoom and inclination in the plane
D : Head pitch and yaw

Figure 1: *Example of face tracking. First and second moments provides an ellipse which delimit the face in the image*

An important property for a face tracker is stability. Stability is measured as the variation of the postion and size of the detected pixels of the face when the subject is at average distance from the camera and is not moving. We have calculated variances of the moments in pixels with regard to the size of the image on sequences of 20 seconds when the person's head has a certain pose and is not moving. Results are shown in Table 2. The face tracker can be perturbed when the subject is in profile because of the detction of his neck.

3.3. Normalized Face Imagette

Once the face is delimited with an ellipse in the image, the face is converted into a normalized luminance imagette (see Figure 2). The normalized face image offers several ad-

Table 2: *Stability of the position and the size of the detected face*

Pose	Front	Half-profile	Profile
X Center	0,31 %	1,13 %	3,23 %
Y Center	0,64 %	1,05 %	1,58 %
Width	0,55 %	1,08 %	1,38 %
Height	0,64 %	1,14 %	1,38 %

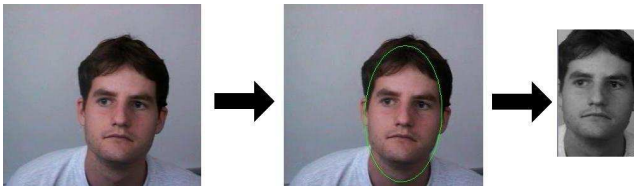

Figure 2: *Face Image Normalization*

vantages : position invariance of face features and scale invariance. This means that processing can be specialized to the normalized scale and processing time is independent of original face size. All further operations take place within this imagette.

4. Generic Face Features Selection

In this section, we search for facial features robust to changes in illumination, pose and identity. We show how to describe an image with receptive fields, then how to automatically learn facial features with clustering and finally determine salient regions of a face.

4.1. Normalized Receptive Fields

Gaussian derivatives provide a feature vector for local appearance that can be made invariant. We use a five dimensional feature vector computed at each pixel by computing the convolution with the first derivative of a Gaussian in x and y direction (G_x, G_y) and the second derivatives (G_{xx}, G_{xy} and G_{yy}). We use grey-level image of the face to be robust to chrominance variations of lights (sun, neon lights,...). We do not use the zeroth order Gaussian derivative in order to remain robust to changes in illumination intensity. Derivatives of higher order have been found to contribute little information for detection [4].

The feature vector (G_x,G_y,G_{xx},G_{xy},G_{yy}) describes the local appearance of a neighboorhood and is determined using Gaussian derivatives that are normalized to the characteristic scale at each pixel. An example of feature vector of a pixel can be seen in Figure 3. The characteristic scale at each pixel is determined with the local maximum

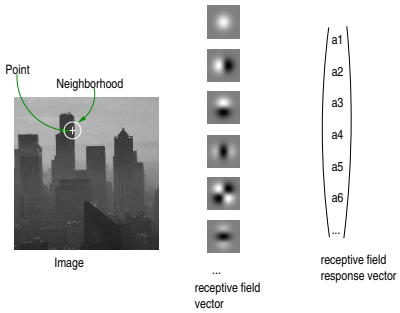

Figure 3: *Appearance based feature vector*

of the Laplacian as function of scale (the scale parameter of the Gaussian), as proposed in [20]. The normalization of face image into an imagette allows us to reduce the range in which the characteristic scale is searched. Two neighboorhoods similar in appearance are close in the feature space. We use a fast, pyramid based, process for determining scale normalized gaussian derivatives [12].

4.2. Clustering operation

K-means clustering is used to determine a combination of Gaussian derivatives that provide a detection of salient facial features that is robust to variations. Gaussian derivatives vectors forms clouds of points in the feature space. The clustering operation finds these clouds. A distance metric for these feature vectors is defined by normalizing feature vectors by their variance. Gathering similar appearances captures the specificity of facial features. Each cluster can be used as a robust detector, because variances are learned. We have found that a single cluster provides a robust detector for salient facial features.

4.3. Robust Facial Features

Applying clustering to the feature vectors for multiple images from several faces provide appearance clusters for background, hair and different skin regions as well as salient facial features. For each pixel, we determine the most probable cluster. A pixel belongs to a certain cluster if the variance normalized distance between the appearance of the pixel and the cluster centroid is minimal in the feature space. Pixels of a same cluster are represented by a point cloud in the feature space and several connected regions in the image. In many experiments, one cluster corresponds to salient facial features and responds to: eyes, nose, mouth and chin.

Detection of facial features can also give rise to a number of small spurious detected regions. These can be eliminated by using a connected components analysis algorithm and compute the bounding box around. Regions with a small

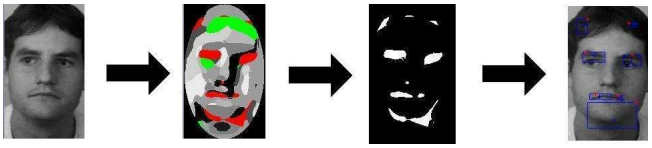

Figure 4: *Robust facial features detection process*
Face image normalization,
Mapping : Regions in red and green are considered as salient robust facial features and reprojected into a binary map,
Connected components analysis

Table 3: *Recall / Precision in % when changing the head orientation in the training process*
Results obtained with a detection threshold of 0.25

Images	Frontal	Near-Frontal	All
Person 1	36,7 / 30,1	40,9 / 4,8	31,1 / 24,6
Person 2	34 / 35,4	35,6 / 4,1	35,2 / 6,7

bounding box are eliminated. The remaining regions correspond to salient facial features. The connected components analysis also gives geometrical informations about robust face features in the face image. Computing the first and second moments of the connected components provides more informations. This information can be reprojected to the original image. The process of robust facial features detection is shown in Figure 4.

5. Experimental Results and Discussion

5.1. Training data

The choice of a good database is crucial for the learning step. To detect salient facial features that are robust under changing conditions, we have used 2 front images of 15 subjects to learn features vectors. Subjects are 20 to 40 years old. Five people have facial hair and 7 people wear glasses. Non-frontal images can introduce noise in the data, because some facial features have different appearances in different poses. Experiments on 2 people are shown in Table 3. Front pose provides more generic appearance for salient facial features, which remain robust on multiple poses after learning, whereas profile images provides appearance for salient features in profile, but not for front.

To remain robust to identity, we have used images of our database, which is composed of series of images of 15 different people. These subjects can be gathered in two classes:

- Class A, in which people's face is "ordinary" or "common" with regard to people in the database. In our database, 73% of the subjects has white skin, european facial type and no beard.
- Class B, in which people's face differs from "common" faces in the database. These people can wear glasses, have a beard or different skin colors. In our database, 27% of the subjects has darker skin or oriental facial type or a beard.

We have observed performances of the learning process with people of different classes and obtained the following results:

- The clustering C(A) is done only with people belonging to class A.
  - Regions obtained for facial features of a subject a $\in$ A are significant and robust.
  - Regions obtained for facial features of a subject b $\in$ B are less significant and more noisy.
- The clustering C(A+a) is done with people belonging to class A and a new subject (a) belonging to class A too. Regions obtained for facial features of the subject (a) are less robust than those obtained with the previous clustering C(A).
- The clustering C(A+b) is done with people belonging to class A and a new subject (b) belonging to class B. Regions obtained for facial features of the subject (b) are less noisy and more salient than those obtained with the previous clustering C(A).

These observations can be explained in the following way. The clustering C(A) done over "common" faces provides better results on subjects of class A, with "common" faces than subjects of class B, whose face's appearance differs from the subjects'faces of class A. Therefore the clustering C(A) is not well adapted for subjects of class B. We must then use other people in our learning process to remain robust to identity.

Adding a new subject a $\in$ A in the clustering does not bring much more information, even on the subject (a). Furthermore, it can lead to a degradation of robustness and more noise, because it will specialize the learning for people of the class A only. Then the class A become more specific, and, as a consequence, regions do not remain generic and robust to identity. This can lead to overfitting.

Alternatively, adding a new subject b $\in$ B provides better detection of facial features on (b), whose appearance differs from those of class A. The clustering C(A+b) adapts to the image of the face of (b) without becoming specialized. Furthermore, salient facial features are more often detected with C(A+b) than with the clustering C(A).

Table 4: *Results obtained with 30 front images and a detection threshold of 0.4*

Number K	2	3	5	7
Recall	11,7 %	22,7 %	70,7 %	30,7 %
Precision	2,3 %	13,1 %	18,2 %	21,5 %
Number K	10	15	20	
Recall	40,2 %	12,2 %	6,1 %	
Precision	47,7 %	57,3 %	11,7 %	

Table 5: *Recall/Precision with regard to detection threshold*

Detection Threshold	0,1	0,25	0,4
Recall	46,3 %	34 %	23,2 %
Precision	22,2 %	25,4 %	26,9 %
Detection Threshold	0,5	0,66	0,75
Recall	17,9 %	10,3 %	7,5 %
Precision	27 %	27 %	27,3 %

5.2. Influence of the number of clusters

The clustering step gathers feature vectors into K clusters. This step is an important part in the learning process and must be carried out. Therefore, the choice of the number of clusters K is crucial. If K is too small, appearance clusters won't be discriminative enough to detect salient features of the face. If K is too big, regions will be too small and too unstable in the image. During our experiments, we tested several K and obtained good results with K = 10. Resulting images with different number of clusters can be seen in Figure 5.

To measure the recall and the precision for each different K, we have employed a 10x15 grid on the normalized imagette of the face (see Table 4). Cases in the grid are manually labelled as follows : 1 if the case contain a facial salient features, 0 otherwise. During the tests, a case of the grid gets the value 1 if the ratio of the number of salient cluster pixels in the case over the total number of pixels in the case exceeds a fixed threshold. This threshold is called the detection threshold (see Table 5).

5.3. Facial feature detection performance

Tests have been made with representative people under changing lighting and pose conditions. The pose is determined by 2 angles (h,v), which vary from -90 degrees to +90 degrees. Each set contains 93 images of the same person at different poses. People are wearing glasses or not and having various skin color. We have calculated the detection rate for each feature for 4 representative people (see Table 6).

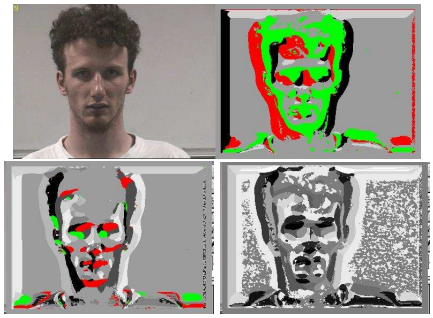

Figure 5: *Influence of the number of clusters*
Regions in red and green are considered as salient robust facial features
Top left image is the original image
Top right image is obtained with 5 clusters, which are not discriminative enough
Bottom left image is obtained with 10 clusters
Bottom right image is obtained with 15 clusters. Regions are too small to be relevant

Table 6: *Facial feature positive detection rate*

Feature	Person 1	Person 2	Person 3	Person 4
Eyes	99 %	97 %	98 %	95 %
Nose	70 %	82 %	61 %	82 %
Mouth	85 %	90 %	95 %	85 %
Chin	84 %	88 %	91 %	84 %
Specificity	-	Glasses	Beard	Matt skin

With an average detection rate of 97 %, eyes are the most often detected feature. Eyes appearance does not vary as much as the other facial features because of their spherical shape. Furthermore, eyes can be detected as blobs on the face under several points of view. Glasses do not affect eyes detection. Mouth detection seems slightly higher for bearded subjects. A mouth is more salient when it is surrounded by a beard.

For 63% of the observed errors, the head pitch is inferior to -30 degrees, so the subject is looking down. This situation represents only 29% of all poses. Indeed, in this situation, eyes are no more visible in the image, but only eyebrows. Therefore, we have trained our algorithm on images on which subjects' head pitch is inferior to -30 degrees. In this case, the resulting clusters are less dicriminating and provides lower detection rate on face images. As a consequence, some facial features, such as chin and eyes, are less salient. Eyes detection is 59% less efficient with the algorithm trained with images of people looking down. The

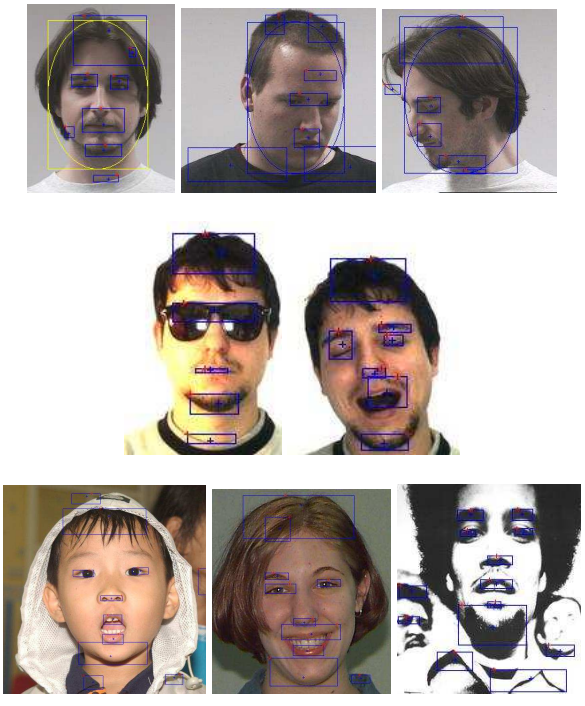

Figure 6: *Examples of facial features detection [18]*

Table 7: *Eyes detection rate with regard to the size of the imagette in pixels*

Size	120x200	120x120	60x100	50x50
Detection	98,2 %	97,8 %	94,2 %	1,8 %

nose has the worst detection average rate with 74%. It does not have as many symmetry properties as eyes and its appearance can suffer many variations. That is why the nose is less often detected than other facial features.

5.4. Influence of the size of the face imagette

To show the importance of the face image normalisation step, we have measured eyes detection rates with different sizes of the face imagette. Results of these experiments can be seen in table 7. Tests have been made on a sequence of 500 images in which the subject moves but has both eyes remaining visible on the screen. The head size changes from 50x50 to 20x20 pixels in the sequences.

The last size, 50x50 pixels, corresponds to face image analysis without normalization, as the face in the sequnce has a maximal size of 50x50 pixels. We can see how the normalisation process enhances the detection rate. This provides the ability to deal with 20x20 pixels images of the head, such as panoramic or wide-angle public cameras images. If this operation is not done, regions will be more imprecise and may not be found. Increasing the size of the normalized face image increases the accuracy of feature detection in the original image of the face. For our experiments, the face imagette has a size of 60x100 pixels.

Conclusions

We have proposed a new approach to detect salient local face features which are robust to pose, illumination and identity. We do not need to constrain the image to allow our algorithm to work. The image is normalized in scale and orientation by a face tracker. Each pixel in the face image is associated to an appearance cluster. One particular cluster stands for salient robust face features which are: eyes, nose, mouth, chin. We have tried to extract and exploit the maximum of informations contained on a single image of a face and to limit the loss of generality.

These regions can be delimited with rectangles in the image. Identifying facial features using positions relative to the face image is difficult because of multiple variations of features possible. These variations are due to changing orientation, emotion and especially identity. Alternatively, a Bayesian classifier should be used to identify the regions. The rectangles can provide a grid on the image of the face too. Robust features can also be used for expression analysis under changing conditions.

References

[1] M. Zobel, A. Gebhard, D. Paulus, J. Denzler, H. Niemann "Robust Facial Feature Localization by Coupled Features" *4th International Conference on Automatic Face and Gesture Recognition*, March 28-30/2000 Grenoble, France pp. 2-7.

[2] D. Hall, J.L. Crowley "Computation of Generic Features for Object Classifi cation" *ScaleSpace*, 2003.

[3] S. J. McKenna, S. Gong, R. P. Wurtz, J. Tanner, D. Banin "Tracking Facial Feature Points with Gabor Wavelets and Shape Models" *1st Int. Conf. on Audio- and Videobased Biometric Person Authentication*, Lecture Notes in Computer Science 1997.

[4] D. Hall "Viewpoint Independant Object Recognition from Local Appearence" *PhD Thesis*, 2001.

[5] K. Schwerdt, J.L. Crowley "Robust face Tracking using Color" *Automatic face and Gesture Recognition*, pp 90-95, 2000.

[6] L. Wiskott, J-M. Fellous, N. Kruger, C. von der Malsburg "Face Recognition by Elastic Bunch Graph Matching" *Pattern Analysis and Machine Intelligence*, Vol 19 pp 775-779, 1997.

[7] J. Malik, S. Belongie, T. Leung, J. Shi "Contour and Texture Analysis for Image Segmentation" *IJCV*, Vol 43 pp 7-27, 2001.

[8] V. Kruger, G. Sommer "Gabor Wavelets Networks for Object Representation and Face Recognition" *Deutsche Arbeitsgemeinschaft fur Mustererkennung*, Vol 22, Sept. 13-15, 2000.

[9] C. Schmid "Constructing Models for Content-Based Image Retrieval" *Computer Vision and Pattern Recognition*, 2001.

[10] D. Lisin, E. Risemann, A. Hanson "Extracting Salient Image Features for Reliable matching using Outlier Detection Techniques" *Computer Vision Systems Third International Conference*, pp 481-491, 2003.

[11] M. Turk, A. Pentland "Eigenfaces for Recognition" *Cognitive Neuroscience*, Vol 3(1), pp 71-96, 1991.

[12] J.L. Crowley, O. Riff "Fast Computation of Scale Normalised Receptive Fields" *International Conference Scale Space*, June 2003, Island of Skye.

[13] Y. Wu, K. Toyama "Wide-Range, Person- and Illumination-Insensitive Head Orientation Estimation" *4th International Conference on Automatic Face and Gesture Recognition*, March 2000 Grenoble, France pp. 183-188.

[14] A.C, Varchmin, R. Rae, H. Ritter "Image based Recognition of gaze Direction using Adaptative methods" *International Gesture Workshop*, Springer, 1997, pp. 245-257.

[15] A.J. Howell, H. Buxton "Active Vision Techniques for Visually Mediated Interaction" *Image and Vision Computing 20(12)*, 2002, pp.861-871.

[16] T. Otsuka, J. Ohya "Real-time Estimation of Head Motion Using weak Perspective Epipolar Geometry" *Proceedings of WACV'98*, October 1998, Princeton.

[17] K.N. Walker, T.F. Cootes, C.J Taylor "Automatically Building Appearance Models from Image Sequences using Salient Features" *IVC(20), No. 5-6*, 15 April 2002, pp. 435-440.

[18] A.M. Martinez and R. Benavente "The AR Face Database" *CVC Technical Report n.24*, June 1998.

[19] G.J. Klinker, S.A. Shafer, T. Kanade "A Physical Approach to Color Image Understanding" *IJCV 1990*

[20] T. Lindeberg "Feature Detection with Automatic Scale Selection" *IJCV 1998 (30) Number 2*, pp. 79-116.