# Matching Deformable Regions Using Local Histograms of Differential Invariants*

Nicolás Pérez de la Blanca[1], José M. Fuertes[2], and Manuel J. Lucena[2]

[1] Department of Computer Science and Artificial Intelligence
ETSII. University of Granada, 18071 Granada, Spain
nicolas@ugr.es
[2] Departmento de Informática. Escuela Politécnica Superior. Universidad de Jaén
Avenida de Madrid 35, 23071 Jaén, Spain
{jmf,mlucena}@ujaen.es

**Abstract.** This paper presents a technique to enable deformable regions to be matched using image databases based on the information provided by the differential invariants of local histograms for the key-region. We shall show how this technique is robust enough to deal with local deformations, viewpoint changes, lighting changes, large motions of the tracked object and small changes in image rotation and scale. The proposed algorithm is based on the building of a specific template where an orthogonal representation space is associated with each of its locations. This space is calculated from neighboring information provided by a vector of local invariants calculated on each of the image's pixels. Unlike other well-known color-based techniques, this algorithm only uses the pixels' gray level values.

## 1 Introduction

In this paper, we shall explore the problem of matching deformable image regions using image databases or image sequences. The basic information used in our work is provided by local histograms of a finite set of image-bands defined from invariant values calculated on the image. What is new about our approach is the template definition which provides us with a very robust approach for dealing with local shape and lighting deformations. Deformable object matching remains a very challenging problem mainly due to the absence of good templates and similarity measures which are robust enough to handle all the geometrical and lighting deformations that might be present in a matching process.

The use of invariant features to match or index objects from images is a well-known approach in computer vision although originally, this was mainly used to characterize objects from their outline shape [11]. In order to recognize objects from their pixel values, different geometrical and lighting differential invariants have been suggested [5],[16],[18]. In practice, however, this type of invariant has only proved useful when applied on points with rich geometrical structures in their neighborhoods [9],[12]. In [6] and [7], a new type of image is introduced where each pixel has an

* This work is partially supported by Grant TIC2001-3316 from the Spanish Ministry of Science and Technology.

J.S. Marques et al. (Eds.): IbPRIA 2005, LNCS 3522, pp. 251–258, 2005.

associated histogram of values rather than a scalar value. This new image concept is the inspiration for our approach, and we shall use it to associate to each pixel a summary of the information defining its neighborhood. In our approach, local histograms obtained after applying each invariant on all image pixels are used as the local features characterizing the neighboring region of each pixel. Our approach is region-based since spatial features better model the type of application we are interested in. Let us consider facial region matching. In recent years, object recognition by parts has been suggested as a very efficient approach for recognizing deformable objects [1],[3],[4]. Although different approaches are used in the recognition process from basic features, the selection and detection of good features is a common task shared by all approaches.

The use of histograms as features of interest can be traced back to Swain & Ballard [17] who demonstrated that color histograms could be used as a robust and efficient mechanism for indexing images in databases. Histograms have been used widely in object and texture recognition and image and video retrieval in visual databases [2], [3],[14]. The main drawback of using global histograms as the main feature is the loss of spatial information. Recent approaches based on the space-scale theory have incorporated the image's spatial information. In [14], multidimensional histograms obtained by applying Gaussian derivative filters to the image are used. This approach incorporates the image's spatial information with global histograms. In [2], while spatial information is also taken into account, a set of intensity histograms are used at multiple resolutions. In [8], it is shown how extremely relevant information for detecting salient regions in the image can be extracted from local histograms at different scales. None of these approaches, however, explicitly addresses the use of the local spatial invariant information present in the image.

In this paper, unlike the approaches mentioned above, we shall attempt to achieve a better compromise between spatial information and robustness to deformations. In our case, the matching template for each image region is built as a spatial array, and a set of histograms (calculated from a spatial neighborhood centered on this position) is associated to each of its positions. Each of these histograms defines a new axis of the representation space associated to each pixel. Building a new orthogonal representation of this space and extracting only the most relevant axis a new parsimonious orthogonal representation of it can be obtained. The projection of the histograms into the new orthogonal subspace provides the coefficient vector used in the matching process. On each image, the template is iterated on all the possible locations within it. The matching score on each image location is the Euclidean norm of the vector difference between the projection coefficients associated to the image and the template, respectively.

This paper is organized into five sections: Section 2 presents the template definition and the matching process; Section 3 presents the gray value invariants we have used in the experiments; Section 4 shows the experimental results; and finally, Section 5 details the discussion and conclusions.

## 2 Template Definition and the Matching Process

Let $\mathcal{R}$ be a region of an image $I$, defining our region of interest (ROI). Different gray level invariants can be calculated on each $\mathcal{R}$ location according to the geometrical or

lighting transformation groups that we expect to deform the region. Standard template matching techniques based on these invariants, however, need only be shown to be effective if applied on image location with a rich gray level structure such as that given by a corner [9][16]. The technique we introduce to characterize the ROI uses a different approach.

Let *ni* be the number of different independent invariants to be calculated on each $\mathcal{R}$ pixel location. Let $\boldsymbol{IB}(\mathcal{R}) = \{\boldsymbol{IB}_1, \boldsymbol{IB}_2, \ldots, \boldsymbol{IB}_{ni}\}$ be the set of band-images calculated by applying each invariant to the region $\mathcal{R}$. An *nbin*×*ni* matrix, $\mathbf{H_I}$, is associated to each pixel location of our ROI where the columns of this matrix are the local histogram in a neighborhood of the pixel from each of the ***IB*** matrices. The bin number, *nbin*, is fixed beforehand and all the histograms are normalized to this value. Each histogram is calculated from a fixed size neighborhood around the pixel.

The set of histograms associated to a pixel can be considered as the different axes of a space characterizing the pixel neighborhood information. According to the gray level structure around the pixel, some of the invariant values provide more relevant information than others. In order to obtain an orthogonal parsimonious representation of this space, we calculate the singular value decomposition on the $\mathbf{H}_I$ matrix, $\mathbf{H_I}=\mathbf{UDV}^T$ and we select the s columns $\mathbf{U}_s=\{\mathbf{U}_1,\ldots,\mathbf{U}_s\}$ associated to the s highest singular values as the new axis of the space. A threshold on the normalized singular values ratio is used to select the most significant ones. The projection of the $\mathbf{H}_I$ matrix into this new space $\mathbf{U}_s$ is given by:

$$\mathbf{c}(x,y) = \mathbf{U}_s^{\mathrm{T}}(x,y) \cdot \mathbf{H}_I(x,y) \tag{1}$$

The $\mathbf{c}(\boldsymbol{x},\boldsymbol{y})$ matrix provides us with the set of coefficients characterizing the pixel location $(\boldsymbol{x},\boldsymbol{y})$. In the matching process, we start by calculating $\mathbf{H_I}$ on each pixel location $(\boldsymbol{r},\boldsymbol{s})$ of the target image. We then calculate a similarity measure on each $(\boldsymbol{r},\boldsymbol{s})$ location by shifting the image template on the target image. The similarity measure is given by:

$$\mathbf{S}(r,s) = \left\{ \sum_{x,y} \left\| \boldsymbol{c}(x,y) - \mathbf{U}_s^{\mathrm{T}}(x,y)\mathbf{H'}_I(x+r, y+s) \right\| \right\} \tag{2}$$

where the sum is on all pixel locations $(\boldsymbol{x},\boldsymbol{y})$ of the region-template. The matrices $\mathbf{c}(\boldsymbol{x},\boldsymbol{y})$ and $\mathbf{U}_s^{\mathrm{T}}(x,y)$ correspond to the template location $(\boldsymbol{x},\boldsymbol{y})$ and the matrix $\mathbf{H'}_I$ to the target image in location $(\boldsymbol{x}+\boldsymbol{r},\boldsymbol{y}+\boldsymbol{s})$. The estimated target location is given by the location of the minimum value of **S** and we use the Euclidean norm.

In our case, all the local histograms are very sparse since the range of gray levels present in the neighborhood of each pixel is usually very small in comparison with the full range of the image. One important consequence of this situation is the need to quantize the image's gray level range before the similarity distances are calculated. A consequence of the quantization process is the invariance to illumination changes which are smaller than the bin width. In all of our experiments, we use a uniform quantization criterion fixing the same length to the interval of the gray levels assigned to each bin. The same process is applied to the gray levels of the template region.

## 3 Gray Level Invariants

In this paper, we use the set of invariants suggested by Schmid in [16]. We only use differential invariants based on the three first order derivatives of the image. The following table shows the invariants used in our experimentation in tensor notation:

$$\nu_S(1,\cdots,8)=\begin{bmatrix} L_l L_l \\ L_l L_{lj} L_j \\ L_{ii} \\ L_{ij} L_{ji} \\ \varepsilon_{ij}(L_{jkl} L_l L_k L_l - L_{jkk} L_l L_l L_l) \\ L_{ij} L_j L_k L_k - L_{ijk} L_l L_j L_k \\ -\varepsilon_{ij} L_{jkl} L_l L_k L_l \\ L_{ijk} L_l L_j L_k \end{bmatrix}, \quad \nu_L(1,\cdots,3)=\begin{bmatrix} \frac{L_l L_{lj} L_j}{(L_l L_l)^{3/2}} \\ \frac{L_{ii}}{(L_l L_l)^{1/2}} \\ \frac{L_{ij} L_{ji}}{L_l L_l} \end{bmatrix}, \quad \nu_L(4,\cdots,7)=\frac{1}{(L_l L_l)^2}\begin{bmatrix} \nu_S(5) \\ \nu_S(6) \\ \nu_S(7) \\ \nu_S(8) \end{bmatrix}$$

$$\nu_L(8)=\frac{\varepsilon_{ij}\varepsilon_{kl} L_l L_{jk} L_l}{(L_m L_m)^{3/2}}, \quad \nu_L(9)=\frac{\varepsilon_{ij} L_j L_k L_{lkl}}{(L_m L_m)^{3/2}} \tag{3}$$

where $\nu_S$ represents the differential invariants associated to the SO(2) similarity group, $\nu_L(1{:}7)$ represents the associated invariant to gray level affine transformations, and $\nu_L(8{:}9)$ represents two invariants associated to lighting reversible transformation [5]. The Cartesian expression of the invariants can be obtained using the usual conventions:

$$L_x=\frac{\partial L}{\partial x},\ L_i=\sum_i L_i = L_x + L_y,\ L_{ij}=\sum_{i,j} L_{ij} = L_{xx}+L_{xy}+L_{yx}+L_{yy}$$
$$\varepsilon_{11}=\varepsilon_{22}=0, \varepsilon_{12}=-\varepsilon_{21}=1 \tag{4}$$

## 4 The Algorithm

The previous steps can be summarized as follows:

1.- Fix the scale value for the histograms.
2.- Fix the set of invariants to be used and calculate their associated image-bands.
3.- Calculate the local-histogram matrix on each location of the template region.
4.- Build up the template $\mathcal{T}(\mathcal{R}_T)$ of the region template using SVD on each of the local-histogram matrices.
5.- For each target image:
  5.1 Build the local-histogram matrix on each location of the image.
  5.2 Shift the template frame on all possible image locations. On each location to project the local-histogram matrices on the orthogonal spaces of the corresponding template location to calculate the image coefficients $\mathbf{c}(\boldsymbol{x},\boldsymbol{y})$.
  5.3 Calculate the similarity measure associated to each template position using (2).
  5.4 Take the image location with the $\mathbf{S}(\boldsymbol{x},\boldsymbol{y})$ minimum value as the best target location.

In order to increase the efficiency of the algorithm, it is applied to a sub-sampled version of the region template and images. From this, we estimate a set of possible points instead of a single location. All these points and their neighboring points (for a fixed size) define the set of points on which we shall apply the algorithm on the original images. The most costly step in this algorithm is the calculation of the similarity maps on each image location. In this respect and taking into account the redundant information present in the template, the error measure given in (2) can only be calculated on a subset of the pixel.

## 5 Experimental Results

Multiple experiments have been performed in order to assess the effectiveness of the proposed algorithm. Firstly, we have focused our experiments on showing how robust our algorithm is to drastic changes in object pose. Secondly, we have also shown how the algorithm is capable of a reasonable level of shape generalization, since with only one sample it is possible to successfully match different instances of the same kind of object. Thirdly, we have shown how robust our algorithm is when there is a very large change in pose and a very hard noise condition. In all of the experiments, we have used a frame with a seven-pixel radius for the histogram estimation. We also quantify the entire histogram range to 32 bins. The active range of the invariant images is selected using a saturation threshold on the invariant values. In our case, a range of values between 100 and –100 was used. In all the experiments, the template region is a rectangular sub-image. In all the experiments, we have tried with different sampling steps (0-4) on the image axis in order to calculate the expression in (2). In all the images, a sampling step of 4 pixels in both axes was sufficient to obtain the highest saliency value in the best location. The full set of the 17 differential invariants has been used in all the experiments.

Video sequences of human heads in motion and two sequences obtained from the Oxford face database[1] have been used in our experiments. Our recorded sequences have one-hundredth images The head in motion sequences were captured in 640x480 format by the same digital camera, but in different lighting conditions. For reasons of efficiency in our experiments, we reduce the image size to the head zone giving 176x224 size images. The Oxford Groundhog-Day database comprises 243 images, which we split into two different sets with men's and women's faces, respectively. The pictures from the Oxford database are 81x81 pixels. Our aim is to match the eyes and the mouth throughout the entire sequence. In our case, the template region was an instance of the matched object chosen from an image of the sequence. However, we also show the results of using a fixed template region on a different image sequence.

In the different rows in Figure 1, we show relevant results for three different sequences where the goal is to match the eye region. The image template for each row is shown in the first cell of the row. The first row shows a person moving their head from right to left as they change their facial expression. The second and third rows show results from the Oxford face database. Figure 2 shows relevant results from the mouth matching experiments. As in Figure 1, the first row shows an image from a

[1] http://www.robots.ox.ac.uk/~vgg/data4.html

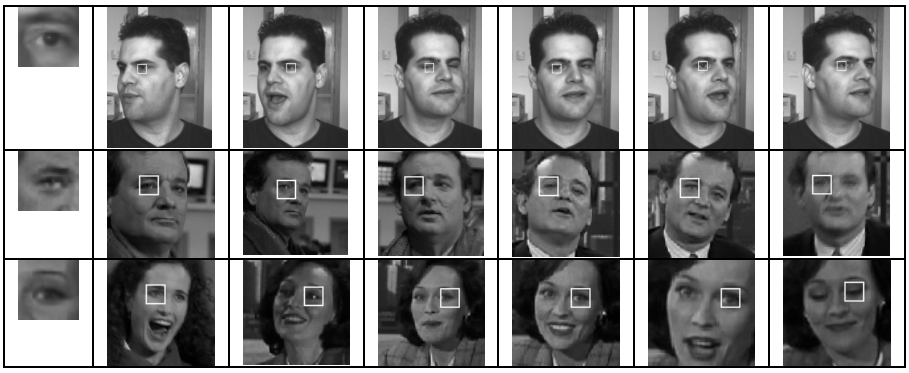

**Fig. 1.** In this figure, the results of the eye-matching problem are shown. In each row, the region-template used is shown in the first column. The white rectangle indicates the best matching region.

**Fig. 2.** This figure shows relevant results for the mouth-matching problem. In each row, the region-template used is shown in the first column. The first row shows images from a recorded sequence. The two last rows show results from the Oxford database. The white rectangle indicates the best matching region.

recorded sequence and the second and third row show the results from the Oxford database.

The experiments show how our algorithm is stable and robust enough for viewpoint changes, local deformations, moderate scale changes and illumination changes. The images in both figures show how our template is flexible enough to match very different instances of an object. This means that the template definition is capable of codifying the relevant information about the object by removing local spatial details. It is also important to emphasize that the algorithm in our experiment is over 90% efficient when the template region and the images are from the same person, but when we match a region template from one person with images from another person, efficiency drops to between 50%-60%. This indicates a lack of generalization that could be explained by the set of used invariants. It is also relevant to point out that the presented results have been obtained when the template-regions cover not only the particular feature of interest but also part of its surrounding area.

In all the experiments, we have only considered translation motions of the template since we are interested in showing that the proposed algorithm is capable of successfully matching a large set of different instances of the original template. Of course, the inclusion of motions such as rotation or scale should greatly improve the technique. One of the main drawbacks of our algorithm is the loss of the image-plane rotation invariance that is present when the full image histogram is considered.

## 6 Conclusions

In conclusion, we have proposed a new matching algorithm for the case of deformable regions and shown its application to face region matching. This algorithm enables us to match different instances of the same object by making use of the information provided by a set of geometrical and lighting invariants. The loss of local order imposed by the use of local histograms has resulted in a high level of robustness in template matching with strong shape deformations even in high noise conditions and moderate lighting changes. Although in theory the algorithm is not robust enough for image-plane rotation and scale, experiments have also shown that there is invariance to small rotations and scale. Full invariance to scale could be obtained by applying a space-scale approach. This, together with achieving higher invariance to lighting changes, shall be one of our future lines of research.

## References

1. S. Agarwal and D. Roth, Learning a sparse representation for object detection, ECCV'02,113-130,2002
2. E.Hadjidemetriou, M.D. Grossberg and S.K. Nayar: Spatial information in multiresolution histograms, In Intern. Conf. CVPR'01, 2001.
3. B.Heisele, P.Ho, J.Wu and T.Poggio, Face recognition: component-based versus global approaches, Computer Vision and Image Understanding 91,6-21,2003
4. R.Fergus, P. Perona and A. Zisserman: Object class recognition by unsupervised scale-invariant learning. In IEEE CVPR'03, 264-271, 2003.
5. L.M.J. Florack, B.M. ter Haar Romeny, J.J. Koenderink and M.A. Viergever, Scale and the differential structure of images, Image and Vision Computing, vol-10, 6, 1992.
6. L.D.Griffin, Scale-imprecission space, Image and Vision Computing 15, 369-398, 1999.
7. J.J.Koenderink and A.J. Van Doorn: The Structure of locally orderless images, Intern. Journal of Computer Vision 31 (273), 159-168, 1999.
8. T. Kadir and M. Brady: Scale, saliency and image description. Intern. Journal of Computer Vision, 45 (2):83-105, 2001.
9. D.G.Lowe, Object recognition from local scale-invariant features. In ICCV'99,1150-1157.
10. J.Matas, O.Chum, M.Urban and T.Pajdla: Robust wide baseline stereo from maximally stable extremal regions. In BMCV'02 Conference, 384-393, 2002.
11. J.L. Mundy and A. Zisserman (eds), Geometric invariance in computer vision, MIT Press, 1992.
12. K.Mikolajczyk and C. Schmid: An affine invariant interest point detector. In ECCV'02, 128-142, 2002
13. W. Niblack. The QBIC project: Querying images by content using color, texture and shape. In Proc. Of SPIE Conf. on Storage and Retrieval for image and video database, vol-1908, 173-187, 1993.

14. B. Schiele and J. L. Crowley. Object recognition using multidimensional receptive field histograms. In ECCV'96, Vol I, pages 610--619, 1996.
15. B. Schiele and J. L. Crowley: Robustness of object recognition to view point changes using multidimensional receptive fields histograms. ECIS-VAP, 1996.
16. C. Schmid and R. Mohr, Local greyvalue invariants for image retrieval, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol 19-5, 530-535, 1997.,
17. M.J. Swain and D.H. Ballard. Color Indexing Intern. Journal of Computer Vision, 7(1):11-32.1991.
18. B.M. ter Haar Romeny, L.M.J. Florack, A.H. Saldem and M.A. Viergever, Higher order differential structure of images, Image and Vision Computing, vol-12, 6, 1994.
19. T.Tuytelaars and L. Van Gool: Wide baseline stereo based on local affinely invariant regions, In British Machine Vision Conference, Bristol, U.K.,412-422. 2000