

Matching Deformable Features Based on Oriented Multi-scale Filter Banks

Manuel J. Marín-Jiménez¹ and Nicolás Pérez de la Blanca¹

Dpt. Computer Science and Artificial Intelligence, University of Granada
C/ Periodista Daniel Saucedo Aranda s/n,
Granada, 18071, Spain
mjmarin@decsai.ugr.es, nicolas@ugr.es

Abstract. This paper presents a technique to enable deformable objects to be matched throughout video sequences based on the information provided by multi-scale Gaussian derivative filter banks. We show that this technique is robust enough for viewpoint changes, lighting changes, large motions of the matched object and small changes in rotation and scale. Unlike other well-known color-based techniques, this technique only uses the gray level values of the image. The proposed algorithm is mainly based on the definition of a particular multi-scale template model and a similarity measure for template matching. The matching approach has been tested on video sequences acquired with a conventional webcam showing a promising behavior with this kind of low-quality images.

1 Introduction

In this paper, we approach the problem of matching deformable objects through video sequences, based on the information provided by responses of oriented multi-scale filter banks. Our approach is traditional in the sense that we define a template of the object of interest, and we attempt to find the image region that best matches the template. What is new about our approach is the template definition and the similarity measure. Deformable object matching/tracking 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 can be present in a matching process. Very recently, object recognition by parts has been suggested as a very efficient approach to recognize deformable object [1][3][2][6]. Different approaches are used in the recognition process from the basic parts, but the matching of salient parts is a common task to all approaches. Region and contour information are the main sources of information from which the location of a part of an object in an image can be estimated (e.g. [7][8]). It has been shown that histograms are robust features for translation, rotation and view point changes [12][10]. However, the main drawback of using the histogram directly as the main feature is the loss of the gray level spatial information [12]. Recent approaches based on the space-scale theory have incorporated the image's spatial information. In [10]

multidimensional histograms, which are obtained by applying Gaussian derivative filters to the image, are used. This approach incorporates the image’s spatial information with global histograms. None of the above approaches explicitly addresses the local spatial information present in the image. The ideas presented in [4] suggest the interest in removing the local spatial information in deformable regions matching process. On the other hand, it is well known that local features based on Gaussian derivatives responses are robust in the task of object recognition [15][5][14], and texture description [13]. In this paper, in contrast to the above approaches based on histograms, we impose a better compromise between spatial information and robustness to deformations. In our case, the matching template for each image region is built as a array of combined responses of oriented multi-scale filter banks. On each image, the template is iterated on all the possible locations within it. The matching on each image location is the vector of the similarity matched on each spatial scale. The optimum (minimum or maximum, according to the similarity criterion definition) of this vector defines the saliency value in each image location. The set of these values defines a saliency map associated to the image, which is the input to the final decision criteria defining the optimum location. This paper is organized in the following way: Section 2 introduces the template definition and the similarity measure; Section 3 presents the algorithm; Section 4 shows the experimental results; and finally, Section 5 concludes the paper.

2 Template and similarity measure

2.1 Template definition

Unlike classical templates based on patches of raw gray levels or templates based on histograms, our approach is based on filter responses. In concrete, the template building is addressed by the HMAX model [9][11]. The main idea is to convolve the image with a filter bank compound by oriented filters at diverse scales. We will use four orientations per scale (0, 45, 90 and 135 degrees).

Let $F_{s,o}$ be a filter bank compound by $(s \cdot o)$ filters grouped into s scales (an even number) with o orientations per scale. Let $F_{i,\cdot}$ be the i -th scale of filter bank $F_{s,o}$ compound by o oriented filters.

The steps for processing an image(or building the template) are the following:

1. Convolve the target image with a filter bank $F_{s,o}$, obtaining a set $S_{s,o}$ of $s \cdot o$ convolved images. The filters must be normalized to zero mean and sum of squares equals one, and also each convolution window of the target image. Hence, values of filtered images will be in $[-1,1]$.
2. For $i = \{1, 3, 5, 7, \dots, s - 1\}$, in pairs $(i, i + 1)$, subsample $S_{i,\cdot}$ and $S_{i+1,\cdot}$ by using a grid of size g_i and selecting the local max value of each grid. Grids are overlapped by v pixels. This is independently done for each orientation. At the end of this step, the resultant images \hat{S}_i and \hat{S}_{i+1} contain the local max values (of each grid) for the o orientations.

- Then, combine each pair \hat{S}_i and \hat{S}_{i+1} in a single band C_i by selecting the max value for each position between both scales $(i, i + 1)$. As a result, $s/2$ bands C_i are obtained, where each one is compound by o elements.



Fig. 1. Anisotropic oriented filters: second order Gaussian derivatives.

The definition of the template can be done in two different ways:

- From the gray-scale image, we can extract (by hand or with an automatic method) a region of interest R , and process R with the previous algorithm. Therefore, each C_i is an independent template.
- On the other hand, we can process an image containing the full model and extract a patch from a C_i band which will be the template. In this case, the template can be extract around a region containing salient points (e.g. global maximums).

We prefer the second option for selecting the template because, as we will see, template matching is carry out in a natural way in the domain of the transformed images $C_{s,o}$.

Note that, by construction, the template is flexible in the sense that it is compound by local maximums, what provides certain invariance to translation, and combination of pairs of scales, what provides some invariance to scale. Moreover, unlike histograms, the template keeps information about local structure. Also, since the template is based on filter responses, invariance to illumination changes is achieved.

The filter bank used in this work is based on second order Gaussian derivatives (as in [13]):

$$G_2(x, y) = \frac{y^2 - \sigma_y^2}{2\pi\sigma_x\sigma_y^5} \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \quad (1)$$

Where σ_x and σ_y are the standard deviations in the directions x and y respectively.

Figure 1 shows a sample of $F_{i,\cdot}$ with its four oriented filters.

2.2 Template matching

Once we have defined our template T , we are interested in locating it in a new image. We will select the position of the new image where the similarity function

raises a maximum. The proposed similarity measure M is based on the following expression:

$$M(\mathbf{T}, \mathbf{X}) = \exp(-\gamma \cdot \|F(\mathbf{T}) - F(\mathbf{X})\|^2) \quad (2)$$

Where \mathbf{T} is the template, \mathbf{X} is the comparison region of the same size of \mathbf{T} , γ controls the steepness of the exponential function, F is an indicator function and $\|\cdot\|$ is the Euclidean norm. Values of M are in the interval $[0, 1]$.

In our first approach F was defined as the identity function (as [11]), but it showed an undesirable behavior due to the influence of the mean. So, $F(\mathbf{X})$ is defined as $F(\mathbf{X}) = \mathbf{X} - \bar{\mathbf{X}}$, where $\bar{\mathbf{X}}$ is the mean of \mathbf{X} . Note that F can become more sophisticated normalizing the energy of the patches or regularizing the standard deviation, however we have empirically checked that it does not worth in the practice.

The problem of find the best matching window \mathbf{X}_p^i with a template \mathbf{T} over an image $C_{s,o}$ can be defined as the optimization function:

$$\mathbf{X}_p^i = \arg \max_{i \in s, p \in N_i} \{M(\mathbf{T}, \mathbf{X}_p^i)\}$$

Where N_i is the set of all possible positions in C_i .

Therefore, the procedure for locating the best matching window is:

1. For all bands C_i and for all possible windows \mathbf{X}_p^i in C_i , compute the similarity measure $M_{i,p}$.
2. Choose the position p at scale i where $M_{i,p}$ is maximum.
3. Transform p to image coordinates.

Since each C_i has been built by subsampling, coordinates p must be transformed to image coordinates to find the actual region in the image where template is found.

3 The algorithm

The previous steps can be summarized as follows:

-
1. Fix the scales for the filter bank $F_{s,o}$.
 2. Build the template \mathbf{T} , following the previously explained method, using $F_{s,o}$.
 3. Transform the target image I with $F_{s,o}$ obtaining $C_{s,o}^I$.
 4. Compute the similarity maps M .
 5. Locate the coordinates p of the global maximum over all positions and bands.
 6. Transform p to image coordinates.
-

When more than a maximum is found, we have decided to choose the position closer to the origin of coordinates. However, other criterions can be defined. For example, if we are working on a video sequence we could choose the position closer to the one in the previous frame.

4 Experimental results

Several experiments have been performed in order to assess the effectiveness of the proposed approach. Firstly, we focus our experiments to show how robust our approach is to diverse perturbations introduced to an object. Secondly, we study the capability of generalization of the templates between different poses and different instances of the objects.

4.1 Parameters for the experiments

For our experiments, the anisotropic second order Gaussian derivatives (with aspect-ratio equals 0.25) are oriented at 0, 45, 90 and 135. All the filter banks contain 8 scales. The standard deviation used for the filter banks is equal to a quarter of the filter-mask size. Table 1 shows the value of the parameters for the filter banks, where FS is the size (in pixels) of the 8 mask-filters and σ is the related standard deviation of the functions. For the subsampling step (see sec. 2.1), grid size g_i is in $\{8, 10, 12, 14\}$ and overlap v is equal to 3.

B	C_1		C_2		C_3		C_4	
FS	7	9	11	13	15	17	19	21
σ	1.75	2.25	2.75	3.25	3.75	4.25	4.75	5.25
GS	8		10		12		14	

Table 1. Parameters of the four bands (B): filter mask size (FS) and filter width (σ) for second order Gaussian derivative filter bank, and grid size (GS) used in subsampling step.

Note that, the value of the grid size will depend on the size of the object we are considering (e.g. faces). So, the value of overlap depends on the minimum value of grid size.

4.2 Measuring robustness

In this section a target image is altered in different ways in order to test the capability of our approach to perform a correct matching in adverse conditions. The experiments has been carried out with functions included in ©*Matlab* 7.0. The six kinds of alterations are:

1. Lighting change: pixel values are raised to an exponent each time.
2. Addition of multiplicative noise (speckle): mean zero and increasing variance in $[0.02:0.07:0.702]$.
3. Blurring: iteratively, a gaussian filter of size 5x5, with mean 0 and variance 1, is applied to the image obtained in the previous iteration.

4. Unsharpening: iteratively, an unsharp filter (for local contrast enhancement) of size 3×3 and α (controls shape of the Laplacian) equals 0.1, is applied to the image obtained in the previous iteration.
5. Motion noise: iteratively, a motion filter (pixels displacement in a fixed direction) with a displacement of 5 pixels in the 45 degrees direction, is applied to the image obtained in the previous iteration.
6. In-plane rotation: several rotations θ are applied to the original image. With values $\theta = [5 : 5 : 50]$.

A template of size 8×8 (with the four orientations) is extracted around the left eye, and the aim is to find its position in the diverse test images. The battery of altered images is shown in figure 2. Each row is compound by ten images. Note that, even for us, some images are really hard.



Fig. 2. The six test. From top to bottom: lighting, speckle, blurred, unsharp, motion, rotation.

In figure 3, we see the similarity maps obtained for the lighting and rotation test. The lightest pixel is the position chosen by our method as the best matching position.

For evaluating the test, the matching is considered correct if the proposed template position is not far from the real one more than 1 unit (in C_i coordinates). The percentages of correct matching for the different cases are shown in table 2.

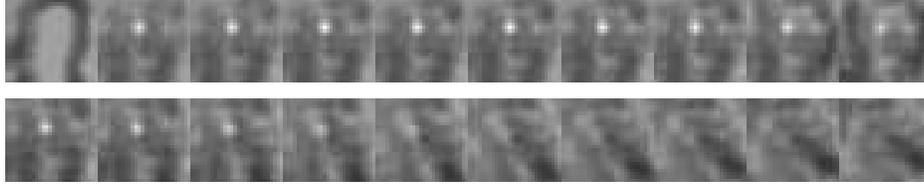


Fig. 3. Responses of similarity measure. Lighter pixels correspond to higher responses. Top row: lighting test. Bottom row: rotation test.

<i>Test</i>	Lighting	Speckle	Blurring	Unsharp	Motion	Rotation
<i>% Hit</i>	90	60	100	100	100	50

Table 2. Percentage of correct matching for each test.

In blurring, unsharpening and motion test the results are really satisfactory, template has been always precisely matched. Matching in lighting test fails only for the first image (left in fig. 2). On the other hand, in speckle test, matching begins failing when variance of noise is greater than 0.5 (the seventh image in the second row, fig. 2); and matching in rotation test fails when angle is near 30 degrees. However, these results suggest the interesting properties of robustness of this kind of templates for matching in adverse noisy conditions.

4.3 Tracking facial features in webcam video sequences

Since nowadays webcams are widely extended and used in diverse environments, and they can be used as input in user-interfaces, in this section we deal with sequences of images taken from a conventional webcam. These sequences contain human heads in motion. Very different poses are present in the images as well as different facial gestures. The size of each frame is 240x320 pixels and they are encoded with ©*Indeo video 5* codec¹. We have converted each frame to gray-scale images and resized to 120x160 pixels for the experiments.

Two templates have been taken from a single frame, and we are interested in locate them in all the remaining frames by using the proposed matching approach. In figure 4 we can see two frames of different subjects and two maps processed at level C_1 and C_2 with a filter oriented at 90 degrees.

The templates shown in figure 5 have been extracted from the first band of subject A (fig. 4), and represents zones around the eye and the mouth. The size of the templates is 6x6 (per 4 orientations). Note that these templates cover a region about 25x25 image pixels (remember the subsampling step in sec. 2.1). They are matched in sequences of more than 150 frames (each one), obtaining

¹ 24 bits color frames, 15 fps, 114 kbps.

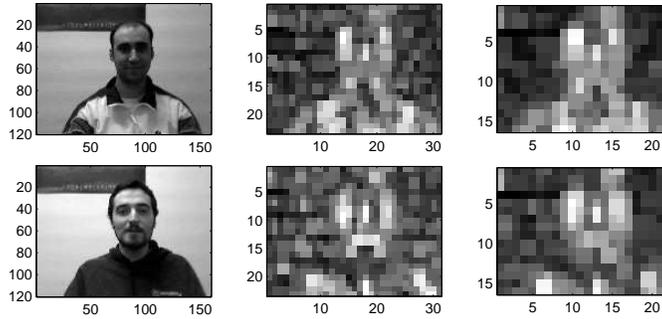


Fig. 4. Webcam images (first column), with C1 maps (from filter oriented at 90) at two scales. Top row: subject A. Bottom row: subject B



Fig. 5. Two templates from subject A (fig. 4). Left: mouth. Right: eye. Size: 6x6. The four orientations are joint for representational purposes.

results as shown in figure 6. Green square and yellow circle refers to region where eye and mouth are matched respectively. The templates not only have been searched for the subject A sequences, but they have been matched in the sequences of subject B. For both subjects the majority of the time matching is correct. It is remarkable the capacity of the templates to generalize between different poses and subjects. However, and what was predictable, figure 7 shows incorrect detections due to large pose changes or big shadows in the regions of interest. In concrete, if face rotates around the Y-axis, it works well² until face is near profile. When face rotates around the Z-axis (in-plane rotation), it works well up to approximately 30 degrees, what supports the results shown in Sec.4.2. And finally, if face rotates around the X-axis (facing up and down), matching performs correctly up to 45 degrees, approximately. Also, there are moments in which the subject approaches to the camera or moves backward, in these situations, where scale changes, matching continues performing well.

Although the quality of the images taken with the webcam is poor, great noise is present and the sequence is compressed the results are promising. Note that none temporal information is used between frames, what can improve the results. Moreover, template is never updated.

² When we say that it works well, we mean that maybe one spurious local maximum appears in some of the intermediate frames.



Fig. 6. Result images with the located position of the two templates (fig. 5). Green square is eye, and yellow circle is mouth. Top row: subject A. Bottom row: subject B.



Fig. 7. Incorrect detection. Result images with the proposed position of the two templates (fig. 5). Green square is eye, and yellow circle is mouth. Top row: subject A. Bottom row: subject B.

5 Summary and conclusions

A scheme for matching deformable regions is proposed and evaluated over human faces. The first experiment shows how our approach is stable and robust enough for different kinds of alterations over the images: changes in illumination, blurring, motion noise, rotation, etc. Then, in the second experiment, templates have been matched along sequences of images taken with a conventional webcam. In these sequences, human faces in motion are present with different points of view and facial gestures. The results show the capability of the approach to match the selected templates in the different frames for the same subject, and for different subjects, showing this way the capability of generalization of the templates. Even though, images acquired with conventional webcams are low-quality, as multiple sources of noise are present, the algorithm as shown quite robust. On the other hand, matching is wrong when too large pose variation is present or huge lighting variation occurs. Nevertheless, this approach is intended to be utilized as tracker initialization or tracker recovery, where temporal information can be used to reduce this problem improving the results. As future work, we intend to compare our proposed scheme with the one based on SIFT features [7].

6 Acknowledgments

Thanks to Dr. Jordi Vitrià for his helpful comments, and to the referees for their suggestions. This work was partially supported by the Spanish Ministry of Education and Science (grant AP2003-2405), and project TIN2005-01665

References

1. Shivani Agarwal, Aatif Awan, and Dan Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE PAMI*, 26(11):1475–1490, Nov. 2004.
2. R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. *IEEE CVPR'03*, 2:264–271, Feb 2003.
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. J. Koenderink and A. van Doorn. The structure of locally orderless images. *Int. Journal of Comp. Vision*, (318273):159–168, 1999.
5. J.J. Koenderink and A.J. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–375, 1987.
6. Bastian Leibe. *Interleaved Object Categorization and Segmentation*. PhD thesis, ETH Zurich, October 2004.
7. David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2(60):91–110, 2004.
8. Krystian Mikolajczyk and Cordelia Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.
9. M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999.
10. Bernt Schiele and James L. Crowley. Object recognition using multidimensional receptive field histograms. In *ECCV (1)*, pages 610–619, 1996.
11. T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *IEEE CSC on CVPR*, June 2005.
12. M.J. Swain and D.H. Ballard. Color indexing. *Int. Journal Comp. Vision*, 1(7):11–32, 1991.
13. M. Varma and A. Zisserman. Unifying statistical texture classification frameworks. *Image and Vision Computing*, 22(14):1175–1183, 2005.
14. J. J. Yokono and T. Poggio. Oriented filters for object recognition: an empirical study. In *Proc. of the Sixth IEEE FGR*, May 2004.
15. Richard A. Young. The gaussian derivative model for spatial vision: I. Retinal mechanisms. *Spatial Vision*, 2(4):273–293, 1987.