Deformable Object Matching Based on Multi-scale Local Histograms

N. Pérez de la Blanca[1], J.M. Fuertes[2], and M. 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 objects to be matched throughout video sequences based on the information provided by the multi-scale local histograms of the images. We shall 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 histogram matching.

1 Introduction

In this paper, we approach the problem of matching deformable objects through video sequences based on the information provided by the gray level histogram of the local neighborhoods of the images. Our approach is traditional in the sense that we shall define a template of the object of interest, and we will 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 [4][5][1]. 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. [17][11][9][10]). Our approach is region-based since gray level features better model the type of application that we are interested in. Let us consider facial region matching. The main features we use are the local histograms at different spatial scales of the image. Furthermore, it is well known that histograms are robust features for translation, rotation and view point changes [14] [15] .

The use of histograms as features of interest can be traced back to Swain & Ballard [15] who demonstrated that color histograms could be used as a robust and efficient

F.J. Perales and B.A. Draper (Eds.): AMDO 2004, LNCS 3179, pp. 154-162, 2004.

mechanism for indexing images in databases. Histograms have been used widely in object and texture recognition and image and video retrieval in visual databases [13] [3] [12]. The main drawback of using the histogram directly as the main feature is the loss of the gray level spatial information [12][15]. Recent approaches based on the space-scale theory have incorporated the image's spatial information. In [13] 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. In [3], spatial information is also taken into account, but using a set of intensity histograms at multiple resolutions. None of the above approaches explicitly addresses the local spatial information present in the image. The ideas presented in [7], [6] suggest the interest in removing the local spatial information in deformable regions matching process. In [8] it is shown that very relevant information to detect salient regions in the image can be extracted from local histograms at different scales.

In this paper, by contrast with the above approaches 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 spatial array, and to each of its positions, the gray level histograms (calculated from a growing sequence of spatial neighborhoods centered on this position) are associated. Although this spatial information is extremely redundant, as we show in the experiment in the case of high noise, this redundancy is extremely useful when estimating the correct location of the template. 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 matching 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 details the discussion and conclusions.

2 Template and Similarity Measure

Let $\mathcal{R}$ be a region of the image I. Let $\mathbf{D}_s(\mathbf{a})=\{\mathbf{x}\in \mathcal{R} \mid \|\mathbf{x}-\mathbf{a}\|<s, \mathbf{a}\in \mathcal{R}, s\in R^+\}$ be the set of points inside the region $\mathcal{R}$ to a distance s of the points $\mathbf{a}$. Let $\mathbf{N}_s=\{\mathbf{D}_s(\mathbf{a}) \mid \mathbf{D}_s(\mathbf{a}) \subset \mathcal{R}\}$ be the set of all local discs $\mathbf{D}_s$ fully contained inside the region $\mathcal{R}$. The set $\{\mathbf{N}_s, s\in \mathcal{S}\}$ represents the information present in the image for the range of scales defined by $\mathcal{S}$. The main drawback of classical template matching methods is the rigidity imposed by the spatial disposition of the gray level values of the pixels, which prevent the template from adapting to the gray level changes on the surface of the object due to lighting or geometrical changes. The template we introduce to characterize a region removes this constraint by associating to each pixel a set of histograms instead of only one gray level. Obviously, small changes in the gray level values around a pixel can be absorbed inside its histogram.

The template associated to a region $\mathcal{R}$ is then defined by the set

$$\mathcal{T}(\mathcal{R}) = \{\mathcal{V}(\mathbf{N}_s), s \in S\} \tag{1}$$

where $\mathcal{V}(\mathbf{N}_s)$ represents the set of histograms built up from the set of spatial neighborhoods $\mathbf{N}_s$.

The next step is to define a similarity measure associated to this template. We have considered the following vector-valued similarity distance between two templates associated to two regions of the same size

$$\mathcal{D}(\mathcal{T}(\mathcal{R}_1), \mathcal{T}(\mathcal{R}_2)) = \min_{s \in S} \left\{ \frac{1}{M_s} \sum_{x \in \mathbf{N}_s} w(x) \left\| \boldsymbol{v}_1(x,s) - \boldsymbol{v}_2(x,s) \right\| \right\} \tag{2}$$

Where $\boldsymbol{v}_i(x,s)$ defines the gray level histogram calculated at the pixel location x and scale value s and M_s is the number pixel in $\mathbf{N}_s$. The values $w(x)$ weight the error norm inversely proportional to the distance of the pixel x to the target region center. In our case we have use the radially symmetric function $w(x)=(1-x)$, if $0 \leq x \leq 1$ and 0 if $x>1$ to define the weights. Different norms calculating a distance between two dense histograms, in matching process, have already been studied [13]. In our case, all the local histograms $\boldsymbol{v}(x,s)$ 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. We have experimented with the Minkowski´s norm (M) for p=1,2 and the capacitory discrimination (C), an entropy-based discrimination measured equivalent

$$\begin{aligned} M_p(\boldsymbol{v}_1, \boldsymbol{v}_2) &= \left\| \boldsymbol{v}_1 - \boldsymbol{v}_2 \right\|_M = \left(\sum_{n \in bin} \left| v_{1n}(x,s) - v_{2n}(x,s) \right|^{\mathrm{p}} \right)^{\frac{1}{\mathrm{p}}}, \mathrm{p} = 1,2 \\ C(\boldsymbol{v}_1, \boldsymbol{v}_2) &= \left\| \boldsymbol{v}_1 - \boldsymbol{v}_2 \right\|_C = 2H\left(\frac{\boldsymbol{v}_1 + \boldsymbol{v}_2}{2}\right) - H(\boldsymbol{v}_1) - H(\boldsymbol{v}_2) \\ H(\mathbf{p}) &= -\sum_{n \in bin} p_n \log p_n \end{aligned} \tag{3}$$

to a simetrized divergence information (see [16]).

One important consequence of the histogram sparseness is the need to quantize the image gray level range before the similarity distances are calculated. It is important to note that in contrast with the results shown in [15], the bin number after the quantization process appears as a relevant parameter. We have tried to estimate this number using the statistical criteria developed for optimum bin number estimation [2], but unfortunately these estimators do not represent, in general, an adequate bin number for the matching process. In our case the bin number used in the experiments was fixed by hand. A consequence of the quantization process is the invariance to illumination differences less than the bin width. In all of our experiments, we use a uniform quantization criterion fixing the length of the interval of the gray levels of the image assigned to each bin. The same process is applied to the gray levels of the template region.

In order to estimate $\mathcal{L}(\mathcal{T}(\mathcal{R}_{\mathcal{T}}),\mathcal{T}(I))$, the set of possible occurrences of the template in an image, we apply the function defined in (2) on each scale and on all the possible image locations in which the template region can be embedded within the image. These values define the saliency vector-map associated to the image I. The set $\mathcal{L}$ is defined by the union of all spatial local minima present on each scale of the saliency map.

$$\mathcal{L}(\mathcal{T}(\mathcal{R}_{\mathcal{T}}),\mathcal{T}(I)) = \bigcup_{s\in S}\left\{ x\in I \;\middle|\; \mathcal{D}_{\mathcal{R}_{\mathcal{T}}}(\mathcal{T}(I_x)) < \mathcal{D}_{\mathcal{R}_{\mathcal{T}}}(\mathcal{T}(I_y)) \;\; \forall y \in \mathbf{D}_{\mathbf{s}}(x) \right\} \quad (4)$$

where $\mathcal{D}_{\mathcal{R}_{\mathcal{T}}}(\mathcal{T}(I_x))$ means $\mathcal{D}(\mathcal{T}(\mathcal{R}_{\mathcal{T}}),\mathcal{T}(I_x))$, and the subscript x indicates the image point where the template is centered. More sophisticated matching techniques can be applied on a subset of these points to decide the best of all of them.

3 The Algorithm

The previous steps can be summarized as follows:

1.- Fix the scale range.
2.- Build up the template $\mathcal{T}(\mathcal{R}_{\mathcal{T}})$ of the region template for the prefixed range of scales.
3.- For each image
  3.1 Build up the template of the i-th image $\mathcal{T}(I_i)$
  3.2 Calculate the saliency map between $\mathcal{T}(\mathcal{R}_{\mathcal{T}})$ and $\mathcal{T}(I_i)$
  3.3 Calculate $\mathcal{L}$, the union of all local minima on all scales for a prefixed neighborhood size.
4.- Apply a final matching criterion on the set of point $\mathcal{L}$.

In order to speed up the efficiency of the algorithm we start applying the algorithm to a sub-sampled version of the region template and images, where the scale values were divided accordingly. In this case, the estimated points and its neighbourhoods for a fixed size define the set of point $\mathcal{L}$. In order to get the maximum accuracy in our matching process the step 4 is carried out on the original images. It is also important for the efficiency in time to implement the histogram calculation using an adaptive process along all the image locations. The most costly step in this algorithm is the saliency map calculation on each image location. In this respect and taking into account the information redundancy present in the template, the error measure given in (2) can only be calculated on a subset of the pixel. In order to remove the produced error by a constant difference in illumination, the template and the target region are fixed to the same average value before calculating the error differences.

4 Experimental Results

Several experiments have been performed in order to assess the effectiveness of the proposed algorithm. Firstly, we focus our experiments to show how robust our algorithm is to drastic changes in the object pose. Secondly, we also show how the

algorithm is capable of a reasonable level of shape generalization, since with only one sample it is possible to successfully track different instances of the same kind of object. Thirdly, we show how robust our algorithm is when there are a very large change in pose and very hard noise condition. In all the experiments, the final decision criterion has been to take the location with the highest saliency measure. In all the experiments, the scale range used was s=2,3,4,5,6, which means they are circular neighborhoods with diameters ranging from 3 to 11 pixels. In all the experiments, the template region is a rectangular sub-image. The background pixels, when present, are removed using a binary mask. The bin number was fixed on each experiment by hand. The three distances between histograms, M_1, M_2 and C, were used in our experiments. Although no very significant differences were detected among them, the Euclidean distance M_2 obtained in all the experiemnts the more accurate matches. All results shown in this section are referred to the Euclidean distance.

We have used video sequences of human heads in motion for the first two experiments, and sequences obtained from the national TV signal for the third experiment. The head in motion sequences were captured in 640x480 format by the same digital camera, but in different lighting conditions. The aim is to match the eyes and the mouth throughout the entire sequences. In our experiments, 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 the generic template region on different image sequence. For reasons of efficiency in our experiments, we reduce the image size to the head zone giving 176x224 size images.

Figure 2 shows the images of a person turning his head from right to left and vice versa. In this case, the region templates for the eyes and mouth, respectively, have been taken from the region of the right eye and the mouth of one of the central images with the head in a front-to-parallel position. Figure 1 shows the two region templates

(a) (b)

Fig. 1. a) Template region used for tracking the eyes; 24x16 pixels, b) Template region used for tracking the mouth; 36x16 pixels. Both template regions belong to the sequence shown in figure 2

Figure 3. shows the results obtained using generic template regions obtained form figure 3 (a) on an image sequence with different expression and very strong changes in the point of view of the face. This experiment shows the robustness of the algorithm for matching very different instances of the same object.

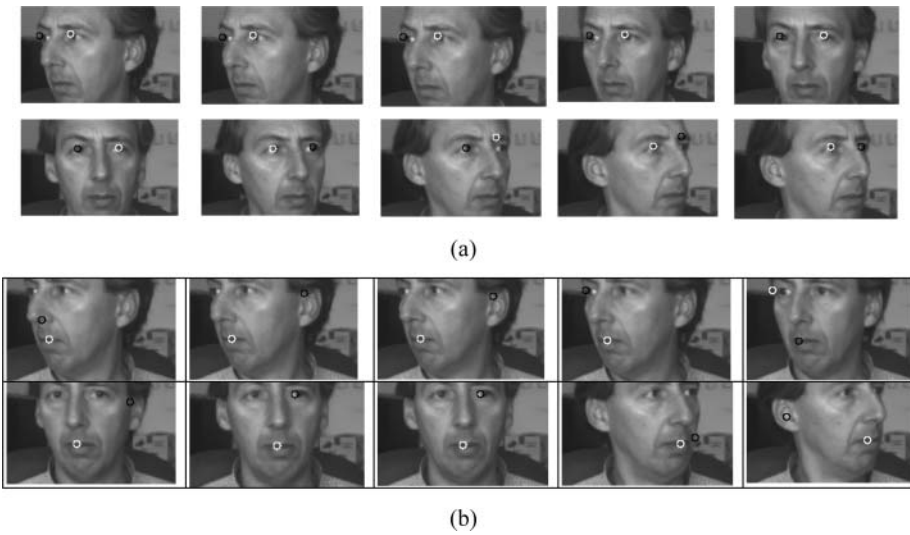


Fig. 2. a) Pieces of relevant images from the eye tracking sequence;. b) Pieces of relevant images from the mouth tracking sequence. The white circle indicates the highest saliency point. The black circle indicates the second highest saliency point.. The shown images are 50x170 pixels

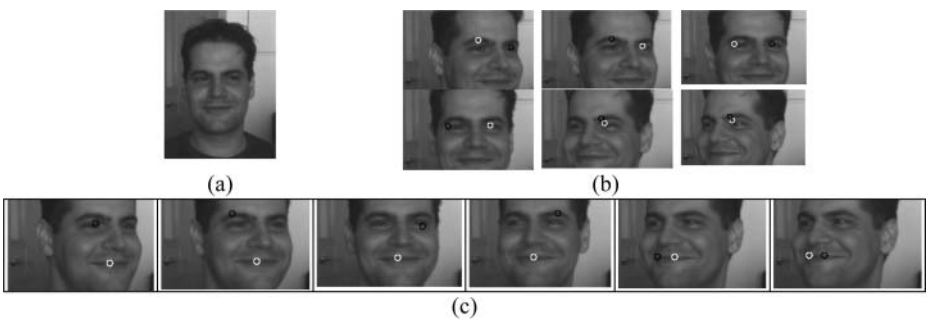


Fig. 3. Robustness of the tracking using generic templates; a) Template image; b) The eye tracking using the eye template extracted from the image (a); c) The mouth tracking using the mouth template extracted from the image (a). The white circle indicates the highest saliency point. The black circle indicates the second highest saliency point. The shown images are 50x170 pixels

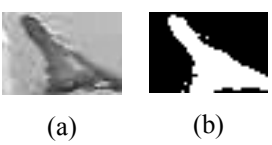
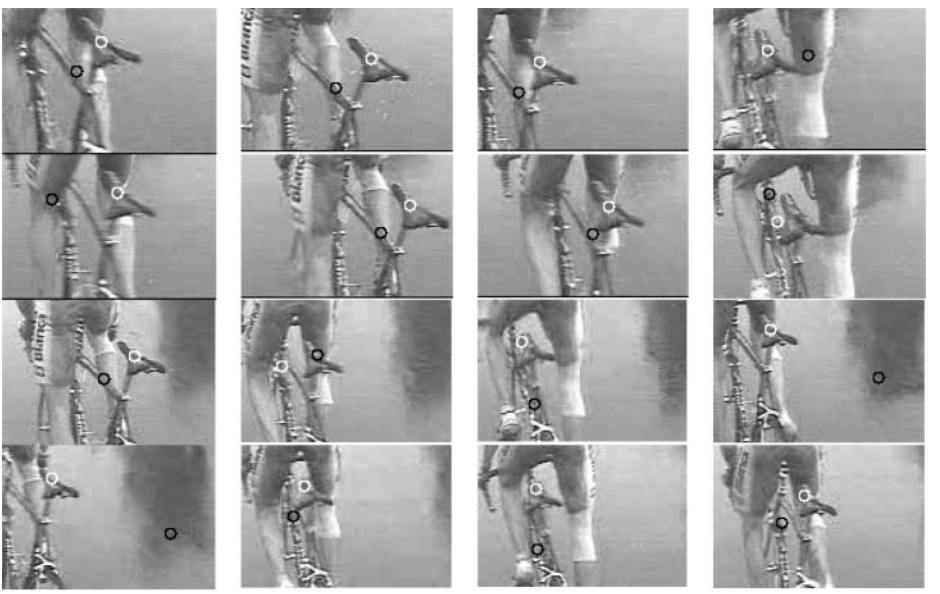


Fig. 4. a) Template region used in the Figure 5. sequence tracking.; 24x16 pixels; b) Binary mask used to remove background pixels

(a)



(b)

Fig. 5. a) Four images of the full sequence are shown; b) A subset of relevant details of the tracking sequence is shown. The white circle indicates the highest saliency point. The black circle indicates the second highest saliency point

In figure 5 we show a sequence recorded in a bicycle race. The aim is to track the bicycle saddle throughout the sequence. Figure 4 shows the template region obtained

from one image in the sequence. In this sequence, the fast bicycle motion joined to the moving camera produces large changes of the saddle viewpoint. In this case, the level of noise is also very high for several reasons: firstly, the digitalization process reduces the images from PAL format to QCIF format; secondly, large lighting changes can be observed throughout the sequence; and thirdly, because of the effect of the rain and fog present in the scenery (see figure 5 (a)).

In all the experiments, we have tried with different sampling steps (0-4) on the image axis in order to calculate the expression (2). In almost all the images a sampling step of 4 pixels in both axes was sufficient to obtain the highest saliency value in the best location. However, in some cases with clutter background or large changes in geometry or lighting, all the pixels had to be considered.

5 Discussion and Conclusions

The first experiment (Figure 2) shows how our algorithm is stable and robust enough for viewpoint changes. The right eye, defining the region template, is always matched as the best location throughout the entire sequence. We also show how the loss of local spatial information has the advantage of matching different instances of the same object but with different shapes. The left eye is the second best location in all the images. Furthermore, Figure 3 also shows 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 removing the local spatial details. In the last experiment (Figure 5), robustness to a non-Gaussian high level of noise and drastic changes in the point of view is shown. It is important to remark that in this difficult sequence of 150 images only in very few images the best location is the second in saliency value.

In all the experiments we have only considered translation motions of the template since our interest is to show that the proposed algorithm is capable of successfully matching a large set of different instance of the original template. Of course the adding of motions as rotation or scale should improved very much 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. The approaches given in [13][3] do not present these problems since they consider full image histograms. However, in preliminary experiments carried out considering global histograms instead of local histograms, poorer results were obtained. In any case, a comparative study of all these techniques would be an interesting future line of research.

In order to compare these results with the traditional correlation matching algorithms we run the same experiments using this algorithm but we obtained completely unsatisfactory results in terms of the object matching position.

In conclusion, the proposed algorithm represents an efficient generalization of the classical matching by correlation algorithm for the case of deformable objects. This algorithm enables us to match different instances of the same object obtained from a very wide viewpoint range. The loss of the local order imposed by the local histogram uses have revealed a high level of robustness in template matching with strong shape deformations even in high noise conditions. It has also proved to be robust enough for

lighting changes by using histograms with a suitable bin number. Although in theory the algorithm is not robust for image-plane rotation and scale, experiments have shown that there is also invariance to small rotations and scale.

Acknowledgement

This work has been financing by the Grant TIC-2001-3316 from the Spanish Ministry of Science and Technology

References

[1] S. Agarwal and D. Roth, Learning a sparse representation for object detection, ECCV'02,113-130,2002

[2] L.Birgé and Y.Rozenholc. How many bins should be put in a regular histogram. Technical Report 721. Laboratoire de Probabilités et Modèles Aléatoires, CNRS-UMR 7599, Université Paris VI & Université Paris VII. 2002.

[3] E.Hadjidemetriou, M.D. Grossberg and S.K. Nayar: Spatial information in multiresolution histograms, In Intern. Conf. CVPR'01, 2001.

[4] 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

[5] R.Fergus, P. Perona and A. Zisserman: Object class recognition by unsupervised scale-invariant learning. In IEEE CVPR'03,264-271,2003.

[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 318273),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] K.Mikolajczyk and C. Schmid: An affine invariant interest point detector. In ECCV'02, 128-142,2002

[12] 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.

[13] B. Schiele and J. L. Crowley. Object recognition using multidimensional receptive field histograms. In ECCV'96, Vol I, pages 610--619, 1996.

[14] B. Schiele and J. L. Crowley: Robustness of object recognition to view point changes using multidimensional receptive fields histograms. ECIS-VAP, 1996.

[15] M.J. Swain and D.H. Ballard. Color Indexing Intern. Journal of Computer Vision, 7(1):11-32.1991.

[16] F.Topsoe. Some inequalities for information divergence and related measures of discrimination. IEEE Trans. Information Theory vol. IT-46, pp. 1602 - 1609, July 2000

[17] 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