Morphological Shape Context: Semi-locality and Robust Matching in Shape Recognition

Mariano Tepper[1], Francisco Gómez[1], Pablo Musé[2], Andrés Almansa[3], and Marta Mejail[1]

[1] Universidad de Buenos Aires, Argentina
[2] Universidad de la República, Uruguay
[3] Telecom ParisTech, France

Abstract. We present a novel shape recognition method based on an algorithm to detect contrasted level lines for extraction, on Shape Context for encoding and on an *a contrario* approach for matching. The contributions naturally lead to a semi-local Shape Context. Results show that this method is able to work in contexts where Shape Context cannot, such as content-based video retrieval.

1 Introduction

The problem of Shape Matching and Recognition can be described as a three-stage process [1]: *(i)* edge detection; *(ii)* invariant coding and matching of semi-local shapes; and *(iii)* grouping of local shape matches.

The first step is most commonly solved by the use of a Canny edge detector which has at least two drawbacks: *(a)* several parameters have to be manually tuned depending on contrast and noise; and *(b)* edges are represented as a non-structured set of edge points which needs to be later grouped into curves, which is a non trivial and error prone task. In this work we substitute the Canny edge detector by a refinement of the Meaningful Boundaries (MB) algorithm [2]. The representation of edges as well-contrasted pieces of level-lines (inspired from mathematical morphology) avoids the edgel linking stage, and the use of Gestalt-inspired [3] *a contrario* detection theory [2] provides a theoretically sound and effective means of selecting parameters and the contrast/noise trade-off. In addition our refinement (see section 3) eliminates the main shortcomming of the basic MB algorithm, thus avoiding that low-contrast parts of the level-lines keep well-contrasted parts from being detected. Fig. 1 compares our MB refinement with the Canny edge detector. Observe that the use of continuous level-lines extracted from a bilinearly interpolated image provide much more finer-grained information, solve the edge-linking problem more effectively and does not introduce a significant computational penalty (thanks to the bilinear FLST [1]).

Once shapes have been extracted from the image, a suitable representation to describe them has to be chosen (step *(ii)* above). Belongie et al. proposed a shape descriptor that is called Shape Context (SC) [4]. SC has many advantages and has been used succesfully in several applications. SC encodes shapes from

(a) (b) (c) (d)

Fig. 1: (a) original image; (b) an area on its upper left corner; (c) detailed view of Canny's filter applied to (a); (d) detailed view of MB applied to (a).

the edge map of an image and it therefore inherits its aforementioned drawbacks. The novel contribution of this work (see section 3) is to fuse SC and MB in what we call Morphological Shape Context (MSC). Results presented further show that this descriptor is able to work in contexts where SC cannot.
The matching step is the least studied of all the processes involved in visual recognition. Most methods use a nearest neighbor approach to match two sets of descriptors [5]. In this work we present an a contrario shape context matching criterion (see [6] and section 4), which gives a clear-cut answer to this issue.

Shape matching as described so far (step *(ii)*) only allows to match relatively simple semi-local shapes. More complex shapes will be represented by groups of shapes that are geometrically arranged in the same manner in both images. Such groups can be detected as a third clustering step. In this work we do not describe this stage in detail but use a basic RANSAC [7] implementation in section 5, in order to experimentally evaluate the results of steps *(i)* and *(ii)* in the context of content-based video retrieval applications.

2 Shape Extraction

This section addresses the problem of extracting the shapes present in an image. We make use of the Fast Level Set Transform (FLST) method where the level sets are extracted from an image, and we propose an extension of the MB algorithm [2], that detects contrasted level lines in grey level images. Let C be a level line of the image u and $x_0, x_1, \ldots, x_{n-1}$ denote n regularly sampled points of C, with geodesic distance two pixels, which in the *a contrario* noise model are assumed to be independent. In particular the gradients at these points are independent random variables. For $x_i \in C$, let μ_j $(0 \leq j \leq n-1)$ be the j-th value of the increasingly sorted vector of the contrast at x_i defined by $|Du|(x_i)$ (the image gradient norm $|Du|$ can be computed on a 2×2 neighborhood).

The curve detection algorithm consists in adequately rejecting the null hypothesis $\mathcal{H}_0$*: the values of* $|Du|$ *are i.i.d., extracted from a noise image with the same gradient histogram as the image* u *itself.*

Following [8], for a given curve, the probability under $\mathcal{H}_0$ that at least k among the n values μ_j are greater than μ is given by the tail of the binomial law

$\mathcal{B}(n, k, H_c(\mu))$, where $H_c(\mu) = P(|Du| > \mu)$. The regularized beta function I can be regarded as an interpolation of the binomial tail to the continuous domain and can be computed much faster. Thus it is interesting, and more convenient, to extend this model to the continuous case using the regularized incomplete beta function $I(H_c(\mu); l_1(k), l_2(k))$, where $l_1(k) = \frac{l}{2}\frac{n-k}{n}$ and $l_2(k) = 1 + \frac{l}{2}\frac{k}{n}$. This represents the probability under $\mathcal{H}_0$ that, for a curve of length l, some parts with total length greater or equal than $l_1(k)$ have a contrast greater than μ.

Definition 1. *Let $\mathcal{C}$ be a finite set of N_{ll} level lines of u. A level line $C \in \mathcal{C}$ is an ε-meaningful boundary if $NFA(C) \equiv N_{ll} \cdot K \cdot \min_{0 \leq k < K} I(H_c(\mu_k); l_1(k), l_2(k)) < \varepsilon$, where K is a parameter of the algorithm. This number is called number of false alarms (NFA) of C.*

As in [1], the expected number of ε-meaningful boundaries in a finite random set of random curves can be proven to be smaller than ε.

Meaningful boundaries usually appear in parallel and redundant groups, because of interpolation. The shape extraction algorithm only detects curves with minimal NFA in such groups [1].

The refinement proposed in Def. 1 is no other than a relaxation of the classic definition by Desolneux *et al.* ([2]) which aims at avoiding underdetection by allowing some parts (up to $k < K$ out of n points) of the curve to be low-contrasted.

The choice of the value of K cannot be directly made as it is highly dependent on the length and the constrast of the curve. Thus the value of K has to be chosen as a function of the curve length and of the image contrast along the curve.

Following Def. 1, we set the value of K as $\hat{K}_\varphi \equiv \arg\max_{i<n} \left(\frac{\sum_{j=0}^{i} \mu_j}{\sum_{j=0}^{n-1} \mu_j} < \varphi \right)$ where $\varphi \in [0, 1]$ is the new parameter of the detection algorithm.

This choice of K is indeed adaptive to the length and contrast of each level line. It is in fact quite stable for values of $\varphi < 0.05$. Larger values lead to an overdetection and, in general, no perceptually significant level lines appear. Studying how this relates with the laws of visual perception is an interesting subject for future research. From a computational and pragmatic point of view, we consider here that this is not a critical parameter that has to be set by the user because: *(i)* all experiments were performed with the same value of $\varphi = 0.02$ obtaining near-optimal performance; and *(ii)* varying the value of φ within the range $(0, 0.05)$ does not significantly affect the results.

3 Shape Encoding

In this section we overview the SC technique [4], and we present an improved version that leads to an intrinsic definition of semi-locality in this new descriptor.

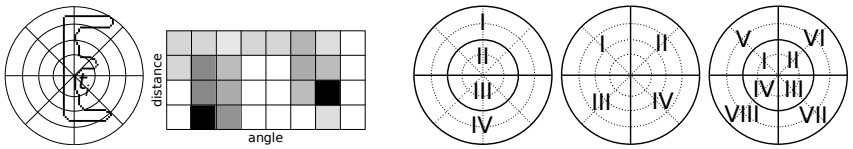
The SC considers a sampled version of the image edge map as the shape to be encoded. The SC of a point in the shape is a coarse histogram of the relative positions of the remaining points. The histogram bins are taken uniformly in log-polar space, making the descriptor more sensitive to positions of nearby sample points than to those farther away.

Let $\mathcal{T} = \{t_1, \ldots, t_n\}$ be the set of points sampled from the edge map of an input image. For each $t_i \in \mathcal{T}$, $1 \leq i \leq n$, the distribution of the $n-1$ remaining points in $\mathcal{T}$ is modeled relative to t_i as a log-polar histogram (Fig. 2a). We denote by $\Theta \times \Delta$ a partition of the log-polar space $[0, 2\pi] \times (0, L]$ into A and B bins respectively, where $L = \max_{t_j \in \mathcal{T}} ||t_j - t_i||_2$. The histogram is defined as

$$SC_{t_i}(\Theta_k, \Delta_m) = \#\{t_j \in \mathcal{T} \ : \ j \neq i, t_j - t_i \in (\Theta_k, \Delta_m)\}$$

where $0 < k \leq A$ and $0 < m \leq B$. The Shape Context of t_i (SC_{t_i}) is defined as a normalized version of $SC_{t_i}(\Theta_k, \Delta_m)$.

Fig. 2a depicts both spatial and matrix representations of a shape context.



(a) (b)

Fig. 2: (a) Shape context of a character 'E'. Left, partition into bins around the point t_i; right, matrix representation of SC_{t_i} (darker means more weigth). (b) Different ways to split a shape context. Doted lines separate bins and thick lines separate bin groupings.

The collection of the SC for every point in the shape is a redundant and powerful descriptor for that shape but has some drawbacks.

First, the sampling stage is performed by considering that the edge map corresponds to a Poisson process [4]. This hard-core model produces a non-deterministic sampling algorithm which means that different runs of the sampling algorithm may give slightly different results. The immediate consequence is that two descriptors from exactly the same image, obtained at different times, may not be equal. In short terms, jitter noise is introduced in the descriptor. In Fig. 3 the effect of the jitter noise is shown, making $d(SC_{t_i}, SC_{t_j}) \approx 0.11 \neq 0$ [4].

Second, from our point of view the main drawback of SC is that it inherits the weaknesses from the edge map. We mentioned previously that extracting curves from the edge map is a hard problem. This fact has a great impact in shape encoding: there is no intrinsic distinction between what is global and what is not. An example is shown in Fig. 3, where $d(SC_{t_i}, SC_{t_k}) \approx 0.3$ which is clearly above the jitter noise $d(SC_{t_i}, SC_{t_j})$. In short terms, a slight modification of the shape has a great impact on the distance. The question "Where does a shape begin and where does it end?" becomes absolutely non trivial. The efforts to overcome this issue lead to heuristic solutions.

[4] $d(\cdot, \cdot)$ is the χ^2 distance and is used throughout this paper

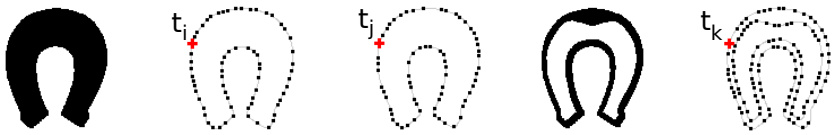


Fig. 3: (a) image horsehoe1; (b) sampled points from horsehoe1; (c) other sampled points from horsehoe1, with the same sampling process than those in (b); (d) image horsehoe2; (e) sampled points from horseshoe2 with the same sampling process than those in (b) and (c). The points t_i, t_j ad t_k are in the same position of the image.

As stated above, the topographic map provides a natural solution to these issues. Meaningful boundaries are much more suitable than the edge map for shape recognition. Meaningful boundaries are used as the set of shapes to be encoded and recognized from an image [1]. Maximal Stable Extremal Regions (MSER), which are very close in spirit to MB, have also been used for shape encoding, see [9] among others.

The main idea is to exploit the benefits of the image structure representation defined in the previous section and to fuse it with SC. We call this new descriptor Morphological Shape Context (MSC).

As in SC, each shape in a given image is composed by a set of points. In MSC, we consider each curve (i.e. meaningful boundary) as a shape. When dealing with curves, the sampling stage is done in a very natural way, by arc-length parameterisation, thus eliminating jitter noise. In the resulting algorithm, shapes are extracted using the MB algorithm. Let us redefine $\mathcal{T} = \{t_1, ..., t_n\}$ as the set of points sampled from a meaningful boundary of an image. The SC is then computed for each sample point t_i, $1 \leq i \leq n$.

Beside the advantages of the representation we described above, one of its keys is the natural separation between level lines (they do not intersect). It allows to go from a global shape encoding to a semi-global one in a natural way, i.e. without fixing any arbitrary threshold. The most powerful advantage is that individual objects present in the image can be matched separately, which was not posible in SC.

In [1] the Level Line Descriptor was designed to detect that two images share *exactly* the same shape. The "perceptual invariance" is only introduced in the matching stage. That is not what we are aiming for. We want to keep the intrinsic "perceptual invariance" given by the SC and be able to detect that two images share two *similar* shapes, independently of the matching algorithm.

4 Shape Matching

As shown in [1], the *a contrario* framework is specially well suited for shape matching. Let $\mathcal{F} = \{F^k | 1 \leq k \leq M\}$ be a database of M shapes. For each shape

$F^k \in \mathcal{F}$ we have a set $\mathcal{T}^k = \{t_j^k | 1 \leq j \leq n_k\}$ where n_k is the number of points in the shape. Let $SC_{t_j^k}$ be the shape context of t_j^k, $1 \leq j \leq n_k$, $1 \leq k \leq M$. As in [6] we assume that each shape context is split in C independent features that we denote $SC_{t_j^k}^{(i)}$ with $1 \leq i \leq C$ (see Fig. 2b for an example).

Let Q be a query shape and q a point of Q. We define $d_j^{k(i)} = d(SC_q^{(i)}, SC_{t_j^k}^{(i)})$. The matching algorithm consists in adequatley rejecting the null hypothesis $\mathcal{H}_0$*: the distances* $d_j^{k(i)}$ *are realizations of* C *independent random variables* $D^{(i)}$, $1 \leq i \leq C$.

Definition 2. *The pair* (q, t_j^k) *is an* ε*-meaningful match in the database* $\mathcal{F}$ *if*

$$\mathrm{NFA}(q, t_j^k) \equiv \left(\sum_{k'=1}^{M} n_{k'} \right) \cdot \prod_{i=1}^{C} P(D^{(i)} \leq d_j^{k(i)} \mid \mathcal{H}_0) < \varepsilon.$$

This number is called number of false alarms (NFA) of the pair (q, t_j^k).

This provides a simple rule to decide whether a single pair (q, t_j^k) does match or not. From one side, this is a clear advantage over other matching methods since we have an individualized assessment for the quality of each possible match. From the other side, the threshold is taken on the probability instead of directly on the distances. Setting a threshold directly on the distances d_j^k (or $d_j^{k(i)}$ for the case) is hard, since distances do not have an absolute meaning. If all the shapes in the database look alike, the threshold should be very restrictive. If they differ significantly from each other, a relaxed threshold would suffice.

Thresholding on the probability is more robust and stable. More stable, since the same threshold is suitable for different database configurations. More robust, since we explicitly control false detections. As proven in [1], the expected number of ε-meaningful matches in a random set of random matches is smaller than ε.

5 Results and Conclusions

In this section we illustrate the performance of the presented methods with three different examples. All the experiments in this paper were produced using $\varphi = 0.02$ for the computation of MB. In both *a contrario* algorithms taking $\varepsilon = 1$ should suffice but we set $\varepsilon = 10^{-10}$ for MB and $\varepsilon = 10^{-2}$ for matching to show the degree of confidence achievable whithout affecting the results.

In the first example, we tested the approach in a video sequence from South Park, which is textureless and composed only by contours. In Fig. 4a, meaningful matches between two consecutive frames are depicted. White dots represent the centers of the MSC. In Fig. 4b, both frames are overlapped to show moving shapes. Note that in Fig. 4a there are no matches in these areas.

The second example, displayed in Fig. 5, is closely related to the first one. Here texture is present and a non-rigid character is moving on the foreground. The matches between frames 3 and 4 of the sequence are shown. Only shapes not

occluded by the movement are matched. The channel logo is correctly matched since it is located in the foreground and it does not move.

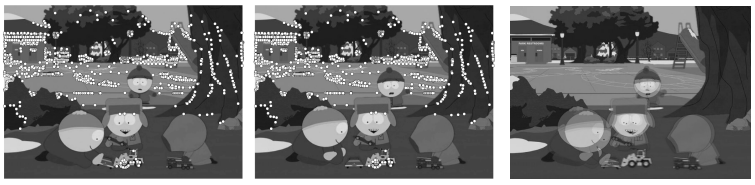
Finally, in Fig. 6 an application to content-based video retrieval is shown. We searched for the parental guidance logo in a video sequence with more than 6000 frames. Fig. 6a depicts the number of matches for each frame of the video. The logo is present in three intervals ([0, 76], [2694, 2772] and [4891, 4969]) which coincide with the three spikes. These spikes are clearly higher than spurious matches in the rest of the video. The second and third spike are smaller than the first one, in those intervals the logo is only at 66% of its original size. This is achieved without any multiscale processing. In Fig. 6b the best match (the correct one) has a NFA of $2.45 \cdot 10^{-9}$ and the worst one (the wrong one), of $9.99 \cdot 10^{-3}$. At $\varepsilon = 10^{-4}$ all matches are correct.

The same experiment as in Fig. 6b using SC gives 3 matches instead of the 29 obtained using MSC (Fig. 6c). All MSC matches are correct and all SC matches are wrong: the global SC approach is unable to match semi-local shapes.

The examples show that semi-locality in the MSC is a key feature to match shapes in contexts where other shapes are present: when very similar images present little differences (Fig. 4), when different foregrounds occlude the same background (Fig. 5), when the query is not present or surrounded by a large set of shapes (Fig. 6). MSC provides a novel approach to deal with such contexts, proving itself successful where SC is not.

References

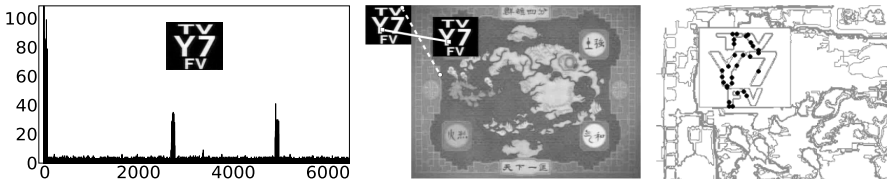
1. Cao, F., Lisani, J.L., Morel, J.M., Musé, P., Sur, F.: A Theory of Shape Identification. Volume 1948 of Lecture Notes in Mathematics. Springer (2008)
2. Desolneux, A., Moisan, L., Morel, J.M.: From Gestalt Theory to Image Analysis. Volume 34. Springer-Verlag (2008)
3. Kanizsa, G.: Organization in Vision: Essays on Gestalt Perception. Praeger (1979)
4. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. PAMI **24**(4) (2002) 509–522
5. Lowe, D.: Distinctive image features from scale-invariant keypoints (2003)
6. Tepper, M., Acevedo, D., Goussies, N., Jacobo, J., Mejail, M.: A decision step for shape context matching. IEEE ICIP (2009) In press.
7. Fischler, M., Bolles, R.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6) (June 1981) 381–395
8. Meinhardt, E., Zacur, E., Frangi, A., Caselles, V.: 3d edge detection by selection of level surface patches. Journal of Mathematical Imaging and Vision
9. Obdrzálek, S., Matas, J.: Object recognition using local affine frames on distinguished regions. In Rosin, P.L., Marshall, D.A., eds.: BMVC. (2002)



(a) (b)

Fig. 4: (a) Matches (white dots) between two frames. There are 1525 matches coherent with a similarity transformation. (b) Both frames overlapped to show moving shapes.

Fig. 5: A video sequence with a non-rigid character moving on the foreground (top). The channel logo is in the bottom right. Matching between frames 6 and 7: there are 141 meaningful matches (white dots) coherent with a similarity transformation.



(a) (b) (c)

Fig. 6: (a) Number of matches per frame of a video with the displayed query. (b) Best (solid line) and worst (dashed line) matches for a target frame. (c) Detail of the logo area with matched points in black dots.