Title
Shape matching and registration by data-driven EM

Authors
Tu, Zhuowen
Zheng, Songfeng
Yuille, Alan

Publication Date
2008-08-21

Peer reviewed

Available online at www.sciencedirect.com
ScienceDirect
Computer Vision and Image Understanding 109 (2008) 290–304
Computer Vision and Image Understanding
www.elsevier.com/locate/cviu

Shape matching and registration by data-driven EM

Zhuowen Tu [a,*], Songfeng Zheng [b], Alan Yuille [c]

[a] Laboratory of Neuro Imaging (LONI), Department of Neurology, UCLA 635 Charles E. Young Drive South, Los Angeles, CA 90095, USA
[b] Department of Statistics, UCLA 8125 Math Sciences Bldg, Los Angeles, CA 90095, USA
[c] Department of Statistics, UCLA 8967 Math Sciences Bldg, Los Angeles, CA 90095, USA

Received 3 May 2006; accepted 18 April 2007
Available online 26 May 2007

Abstract

In this paper, we present an efficient and robust algorithm for shape matching, registration, and detection. The task is to geometrically transform a source shape to fit a target shape. The measure of similarity is defined in terms of the amount of transformation required. The shapes are represented by sparse-point or continuous-contour representations depending on the form of the data. We formulate the problem as probabilistic inference using a generative model and the EM algorithm. But this algorithm has problems with initialization and computing the E-step. To address these problems, we define a data-driven technique (discriminative model) which makes use of shape features. This gives a hybrid algorithm which combines the generative and discriminative models. The resulting algorithm is very fast, due to the effectiveness of shape-features for solving correspondence requiring only a few iterations. We demonstrate the effectiveness of the algorithm by testing it on standard datasets, such as MPEG7, for shape matching and by applying it to a range of matching, registration, and foreground/background segmentation problems.

Keywords: Shape matching; Registration; Soft assign; EM; Shape context

1. Introduction

Shape matching is a very important problem in computer vision. The work in this paper treats shape matching as finding the best geometrical transformation between two shapes in the spirit of Grenander's pattern theory [12].

There is a big literature on the different varieties of shape representations, matching criteria, and algorithms that have been applied to this problem. This includes representations such as Fourier analysis [43], moments [14], scale space [24], and level sets [27]. Some representations, such as FORMS [44] or shock-edit [33], explicitly represent the shape in terms of parts. Other approaches using chamfer matching [35] formulate the problem to enable rapid search through different appearances of the shape. Alternative approaches include manifold learning and metric embedding [15] and hierarchical structures with associative graphs [28]. For more details of shape representations and matching algorithms see recent survey papers [19,39]. Our approach is mostly closely related to the probabilistic formulation using free energies developed by [8,30,21] and the work on shape context [4].

We formulate the problem in terms of probabilistic inference using a generative model. This formulation involves hidden variables which indicate the correspondence between points on the two shapes. This leads naturally to an EM algorithm [9] based on the free energy formulation [26]. Unfortunately computational issues make performing the E-step impractical and also give problems for initialization. We solve these problems by introducing a discriminative model which uses shape features. These shape features are less ambiguous for matching than points, and hence we call them *informative features*. This leads to a hybrid algorithm which combines the generative and discriminative models. See Tu et al. [36]

* Corresponding author.
E-mail addresses: zhuowen.tu@loni.ucla.edu (Z. Tu), sfzheng@stat.ucla.edu (S. Zheng), yuille@stat.ucla.edu (A. Yuille).

1077-3142/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2007.04.004

for more discussion of the trade-off between generative and discriminative methods.

Our approach helps unify existing work by Chui and Rangarajan [8] and Belongie et al. [4]. We use a probabilistic formulation similar to Chui and Rangarajan but combine this with the shape features described by Belongie et al. This leads to an algorithm that has fewer iterations and better initialization than Chui and Rangarajan, while being better at dealing with uncertainty than Belongie et al.

Our approach represents the shapes in terms of points, which are supplemented by shape feature for the discriminative models. We use two types of representations—*sparse-point* and *continuous-contour*. The continuous-contour representation leads to better shape features, since the arc-length is known, but this representation is not always practical to compute. These representations are adequate for this paper, but their lack of ability to represent shape parts makes them unsuitable for matching shapes when parts can be missing (unlike methods such as [44,33]).

The structure of this paper is as follows. Section 2 gives the generative formulation of the problem. In Section 3, we motivate the discriminative approach. Section 4 describes how the algorithm combines the two methods. In Section 5 we give examples on a range of datasets and problems.

2. The generative formulation

The task of shape matching is to match two shapes, X and Y, and to measure the similarity between them. We refer to X as the *target shape* and Y as the *source shape*. We define the similarity measure in terms of the transformation that takes the source shape into the target, see Fig. 1. In this paper we use two types of transformation: (i) a global affine transformation $\mathbf{A}$ and (ii) a smooth non-rigid transformation $\mathbf{f}$.

2.1. Shape representation

We use two types of shape representation in this paper: (I) *sparse-point*, and (II) *continuous-contour*. The choice will depend on the form of the data. Shape matching will be easier if we have a continuous-contour representation because we are able to exploit knowledge of the arc-length to obtain shape features which are less ambiguous for matching, and hence more informative, see Section 3.1. But it may only be possible to compute a sparse-point representation for the target shape (e.g., the target shape may be embedded in an image and an edge detector will usually not output all the points on its boundary).

(a) Target shape X (b) Transformation $(A, \mathbf{f})$ (c) Source shape Y

Fig. 1. An example of shape matching where a source shape Y is matched with a target shape X by a geometric transformation $(A, \mathbf{f})$.

(I) For the **sparse-point** representation, we denote the target and source shape respectively by:

$$X=\{\mathbf{x}_i : i=1,\ldots,M\}, \text{ and } Y=\{\mathbf{y}_a : a=1,\ldots,N\}. \tag{1}$$

(II) For the **continuous-contour** representation, we denote the target and source shape respectively by:

$$X = \{\mathbf{x}(s) : s \in [0,1]\}, \quad \text{and} \quad Y = \{\mathbf{y}(t) : t \in [0,1]\}, \tag{2}$$

where s and t are the normalized arc-length. In this case, each shape is represented by a 2D continuous-contour. By sampling points along the contour we can obtain a sparse-point representation $X=\{\mathbf{x}_i : i=1,\ldots,M\}$, and $Y=\{\mathbf{y}_a : a=1,\ldots,N\}$. But we can exploit the continuous-contour representation to compute additional features that depend on differentiable properties of the contour such as tangent angles.

2.2. The generative model

Our generative model for shape matching defines a probability distribution for generating the target X from the source Y by means of a geometric transformation $(\mathbf{A},\mathbf{f})$. There will be priors $P(\mathbf{A}),P(\mathbf{f})$ on the transformation which will be specified in Section 2.3.

We also define binary-valued correspondence variables $\{V_{ai}\}$ such that $V_{ai}=1$ if point a on the source Y matches point i on the target X. These are treated as hidden variables. There is a prior $P(V)$ which specifies correspondence constraints on the matching (e.g., to constrain that all points on the source Y must be matched).

The choice of the correspondence constraints, as specified in $P(V)$ is very important. They must satisfy a trade-off between the modeling and computational requirements. Constraints that are ideal for modeling purposes can be computationally intractable. The prior $P(V)$ will be given in Section 2.5 and the trade-off discussed.

The full generative model is $P(X,V,\mathbf{A},\mathbf{f}|Y)=P(X|Y,V,\mathbf{A},\mathbf{f})P(\mathbf{A})P(\mathbf{f})P(V)$, where the priors are given in Sections 2.3 and 2.5. The distribution $P(X|Y,V,\mathbf{A},\mathbf{f})$ is given by:

$$P(X|Y,V,\mathbf{A},\mathbf{f}) = \frac{1}{Z}\exp\{-E_G(X,Y,V,\mathbf{A},\mathbf{f})\}, \quad \text{where,}$$
$$E_G(X,Y,V,\mathbf{A},\mathbf{f}) = \sum_{i=1}^{M}\sum_{a=1}^{N} V_{ai}\|\mathbf{x}_i - \mathbf{A}\mathbf{y}_a - \mathbf{f}(\mathbf{y}_a)\|^2. \tag{3}$$

By using the priors $P(\mathbf{A}),P(\mathbf{f}),P(V)$ and summing out the V's, we obtain (this equation defines $E_T[\mathbf{A},\mathbf{f};X,Y]$):

$$P(X,\mathbf{A},\mathbf{f}|Y) = \sum_V P(X|Y,V,\mathbf{A},\mathbf{f})P(V)P(\mathbf{A})P(f) \triangleq \frac{1}{\hat{Z}} \times \exp\{-E_T[\mathbf{A},\mathbf{f};X,Y]\} \tag{4}$$

We define the optimal geometric transformation to be:

$$(\mathbf{A}^*, \mathbf{f}^*) = \arg\max_{\mathbf{A},\mathbf{f}} P(X, \mathbf{A}, \mathbf{f}|Y) = \arg\min_{\mathbf{A},\mathbf{f}} E_T[\mathbf{A}, \mathbf{f}; X, Y]. \tag{5}$$

We define the similarity measure between shapes to be

$$S(X : Y) = E_T[\mathbf{A}^*, \mathbf{f}^*; X, T] = -\log P(X, \mathbf{A}^*, \mathbf{f}^*|Y) - \log \hat{Z}. \tag{6}$$

This is the negative logarithm of probability of generating the target X from the source Y by the optimal transformation, with the normalization constant $\hat{Z}$ removed. This similarity measure will also correspond to the minimum of the free energy, see Eq. (12) (again with the normalization constant removed).

2.3. The geometric prior

The geometric transformation consists of a global (affine) transformation **A** and a smooth non-rigid transformation **f** (which includes a translation term). The prior on the non-rigid transformation **f** will enforce it to be smooth, and will be given at the end of this section. The prior on the affine transformation **A** is defined based on its decomposition into rotation angle θ, scaling S_x,S_y, and shear k. It encourages equal scaling both directions and favors limited shear.

More precisely, we decompose the affine transformation as follows [1]:

$$\mathbf{A} = \begin{pmatrix} S_x & 0 \\ 0 & S_y \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}. \tag{7}$$

The prior on **A** is given by

$$p(\mathbf{A}) \propto \exp\{-E_A(\mathbf{A})\}, \quad \text{and}$$
$$E_A(\mathbf{A}) = E_{\text{scale}}(S_x, S_y) + E_{\text{shear}}(k) + E_{\text{rot}}(\theta), \tag{8}$$

where $E_{\text{scale}}(S_x, S_y) = b_1(\frac{S_x - S_y}{S_x + S_y})^2$, $E_{\text{shear}}(s) = b_2 k^2$, and $E_{\text{rot}}(\theta)$ is constant (i.e., the prior on θ is the uniform distribution). In our experiments, we set $b_1 = 1.0$ and $b_2 = 10.0$. The prior on the non-rigid transformation **f** is of form:

$$p(\mathbf{f}) \propto \exp\{-E_f(\mathbf{f})\}. \tag{9}$$

In this paper, we use the thin-plate-spline (TPS) kernel [5] where we impose that **f** is of form:

$$\mathbf{f}(x, y) = \sum_a \mathbf{w}^a U(x_a - x, y_a - y) + \mathbf{t}, \tag{10}$$

where $\mathbf{w}^a = (w_x^a, w_y^a)$ are vector valued weights, x,y are the components of the vector **y**, $U(x,y) = (x^2 + y^2)\log(x^2 + y^2)$, **t** is translation, $\sum_a \mathbf{w}^a = 0$, and $\sum_a w_x^a x_a = \sum_a w_y^a y_a = 0$. Then we define E_f by:

$$E_f(\mathbf{f}) = \lambda \int\int \left(\frac{\partial^2 \mathbf{f}}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 \mathbf{f}}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 \mathbf{f}}{\partial y^2}\right)^2 \mathrm{d}x\,\mathrm{d}y$$
$$= \lambda \sum_{a,b=1}^{N} \mathbf{w}_a^T U(\mathbf{y}_a - \mathbf{y}_b)\mathbf{w}_b. \tag{11}$$

We also considered an alternative energy function which encourages **f** to be small and smooth (Yuille and Grzywacz [41]). The small-and-smooth energy function usually gave similar results to the thin-plate-spline, but it required us to model the translation **t** separately. We will not discuss it further in this paper. Note that in our ECCV paper [37] our reported results were obtained using the TPS, but we erroneously stated that they were obtained using small-and-smooth.

2.4. The EM algorithm

We use the EM algorithm to estimate $(\mathbf{A}^*,\mathbf{f}^*) = \arg\max_{\mathbf{A},\mathbf{f}} P(X,\mathbf{A},\mathbf{f}|Y)$, where the correspondence variables V are treated as hidden variables which are summed out.

It can be shown [26] that estimating $(\mathbf{A}^*,\mathbf{f}^*)$ from $P(X,\mathbf{A},\mathbf{f}|Y)$ is equivalent to minimizing the EM free energy function:

$$F(Q,\mathbf{A},\mathbf{f}) = -\log P(X,\mathbf{A},\mathbf{f}|Y) + \sum_V Q(V)\log\frac{Q(V)}{P(V|X,Y,\mathbf{A},\mathbf{f})}$$
$$= -\sum_V Q(V)\log\{P(X|Y,V,\mathbf{A},\mathbf{f})P(\mathbf{A})P(\mathbf{f})P(V)\}$$
$$+ \sum_V Q(V)\log Q(V), \tag{12}$$

where $Q(V)$ is a distribution over the correspondence variables. The minimum of the free energy $F(Q,\mathbf{A},\mathbf{f})$ is equal to $\min_{\mathbf{A},\mathbf{f}}\{-\log P(X,\mathbf{A},\mathbf{f}|Y)\}$.

The EM algorithm consists of two steps: (I) The M-step minimizes $F(Q,\mathbf{A},\mathbf{f})$ with respect to $(\mathbf{A},\mathbf{f})$ with $Q(V)$ fixed and will be given in detail in Section 4.(II) The E-step minimizes $F(Q,\mathbf{A},\mathbf{f})$ with respect to $Q(V)$ keeping $(\mathbf{A},\mathbf{f})$ fixed. The E-step at iteration t can be expressed analytically as:

$$Q^t(V) = P(V|X, Y, \mathbf{A}^{t-1}, \mathbf{f}^{t-1}). \tag{13}$$

Unfortunately there are two difficulties with using the EM algorithm in the current form. Firstly, it requires good initialization or it will get stuck in a local maximum of $P(X,\mathbf{A},\mathbf{f}|Y)$. Secondly, the E-step is very difficult to compute unless we put restrictions on the prior $P(V)$ which are so strong that they may degrade the quality of the result, see Section 2.5. These problems motivate us to introduce a discriminative model which is discussed in Section 3.

2.5. The correspondence constraints and $P(V)$

The prior $P(V)$ enforces constraints on the correspondences between points on the target and the source. This leads to a trade-off between modeling and computational requirements.

From the modeling perspective, we would prefer constraints which are *two-sided* between the target and source. These constraints would enforce that most target and source points are matched, and would be flexible enough to allow for missing points. But imposing two-sided constraints makes the E-step of the EM algorithm impractical,

because it is impossible to compute the right-hand side of Eq. (13).

One strategy to deal with this problem is by using a mean field theory approximation together with two-sided constraints. This strategy was used by Chui and Rangarajan [8] and was very successful. We will describe it at the end of this section.

Another strategy is to impose *one-sided constraints* which ensure that all points on the sources are either matched to points on the target, or are unmatched (and pay a penalty). In theory, this would allow situations where a single point of the target is matched to many points on the source, while many other target points are unmatched. In practice, the nature of the geometric transformations can prevent these situations from happening (since they correspond to highly improbable, or impossible, geometric transformations).

In this paper, we choose to impose the one-sided constraint that each point a is either matched to a single point $v(a)$, or it is unmatched and we set $v(a) = 0$. (equivalently $v(a) = \sum_i V_{ai} i$, where $\sum_i V_{ai} = 1$ or 0 for all a). Then we put a prior probability on V, $P(V) \propto e^{-\lambda \sum_{a=1}^{N} (1-v(a))}$, which penalizes missing points. This choice of prior has the major advantage that it is factorizable $P(V) = \prod_a P_a(v(a))$. It means that we can minimize the free energy (12) to solve for:

$$Q(V) = \prod_{a=1}^{N} Q_a(v(a)) = \prod_{a=1}^{N} \frac{1}{Z_a} \exp\{-\lambda - \sum_{i=1}^{M} V_{ai} \|\mathbf{x}_i - \mathbf{A}\mathbf{y}_a - \mathbf{f}(\mathbf{y}_a)\|\}, \quad (14)$$

with $Z_a = \mathrm{e}^{-\lambda} + \sum_{i=1}^{M} \mathrm{e}^{-\|\mathbf{x}_i - \mathbf{A}\mathbf{y}_a - \mathbf{f}(\mathbf{y}_a)\|}$.

This gives an analytic expression for the E-step of the EM algorithm – by replacing Q by Q^t on the left hand side, and $\mathbf{A}$, $\mathbf{f}$ by $\mathbf{A}^{t-1}$, $\mathbf{f}^{t-1}$ on the right hand side of Eq. (14).

We can also use Eq. (14) to obtain a simple expression for the free energy which depends only on $\mathbf{A}$, $\mathbf{f}$. To do this, we first express:

$$P(X|Y,V,\mathbf{A},\mathbf{f})P(\mathbf{A})P(\mathbf{f})P(V) = \frac{1}{Z} \exp\{-E_T[\mathbf{A},\mathbf{f},V;X,Y]\}, \text{ where}$$
$$E_T[\mathbf{A},\mathbf{f},V;X,Y] = E_G(X,Y,V,\mathbf{A},\mathbf{f}) + E_A(\mathbf{A}) + E_f(\mathbf{f}) + \lambda \sum_{a=1}^{N} (1 - v(a))\}. \quad (15)$$

Secondly, we substitute $Q(V)$ from Eq. (14) into the free energy (12) to obtain:

$$\begin{aligned} F(Q,\mathbf{A},\mathbf{f}) &= \sum_V Q(V) E_T[\mathbf{A},\mathbf{f},V;X,Y] + \log \hat{Z} \\ &\quad + \sum_{b=1}^{M} \left\{ \sum_{V_b} Q(V_b) \log Q(V_b) \right\} \\ &= E_A(\mathbf{A}) + E_f(\mathbf{f}) \\ &\quad - \sum_{a=1}^{N} \log \left\{ e^{-\lambda} + \sum_{i=1}^{M} e^{-\|\mathbf{x}_i - \mathbf{A}\mathbf{y}_a - \mathbf{f}(\mathbf{y}_a)\|} \right\} + \log \overline{Z}. \end{aligned} \quad (16)$$

Note that a simplification occurs because $\log Q_b(v(b)) = -\{\lambda V_{b0} + \sum_{i=1}^{M} V_{bi} \|\mathbf{x}_i - \mathbf{A}\mathbf{y}_b - \mathbf{f}(\mathbf{y}_b)\|^2 - \log Z_b$, hence parts of the term $\sum_b Q_b(V_b) \log Q_b(V_b)$ cancels with some terms from $\sum_V Q(V) E_T[\mathbf{A},\mathbf{f},V;X,Y]$. This enables us to rapidly compute the free energy (except for the normalization term $\log \overline{Z}$ which we do not use in our shape similarity criterion).

This E-step in Eq. (14) is computationally simple, but it gives poor performance when applied to the generative model. The number of iterations required for convergence is large (e.g greater than 20) and the algorithm can get stuck in local minima of $F(Q,\mathbf{A},\mathbf{f})$ and produce distorted matching. These errors can be traced to the use of point features for matching and their inherent ambiguity compounded by the use of one-sided constraints.

Our solution is to augment the generative model with a discriminative (data-driven) model, see Section 3. The discriminative model uses shape features which are far less ambiguous for matching, and hence more informative, than point features. The use of shape features was developed by Belongie et al. [4] who used them in conjunction with two-sided constraints requiring one to one correspondence (with some ability to tolerate unmatched points). Note that Belongie et al. did not formulate their approach probabilistically.

An alternative by Chui and Rangarajan [8] makes use of the mean field approximation to the free energy [42,11]. They impose matching constraints $\sum_{a=0}^{M} V_{ai} = 1, \ \forall i$ and $\sum_{i=0}^{N} V_{ai} = 1, \ \forall a$, where the indices $a = 0$ and $i = 0$ denote dummy points which can be used to allow for unmatched points (e.g. $V_{a0} = 1$ implies that point a is unmatched). The prior will add a penalty term to encourage most points to be matched.

A simplified version of Rangarajan and Chui's mean field free energy is expressed in the form:

$$\begin{aligned} F(\{m_{ai}\},\mathbf{A},\mathbf{f}) &= \sum_{i=1}^{M} \sum_{a=1}^{N} m_{ai} \|\mathbf{x}_i - \mathbf{A}\mathbf{y}_a - \mathbf{f}(\mathbf{y}_a)\|^2 \\ &\quad + E_A(\mathbf{A}) + E_f(\mathbf{f}) + T \sum_{i=1}^{M} \sum_{a=1}^{N} m_{ai} \\ &\quad \times \log m_{ai} - \zeta \sum_{i=1}^{M} \sum_{a=1}^{N} m_{ai}, \end{aligned} \quad (17)$$

where $E_f(\mathbf{f})$ is given by Eq. (11), $\mathbf{f}(\mathbf{y}_a)$ is given by Eq. (10), and the variable m_{ai} denotes $Q(V_{ai} = 1)$. By comparison to Eq. (12), we see that the mean field formulation approximates the entropy term $\sum_V Q(V) \log Q(V)$ by $\sum_{i=1}^{M} \sum_{a=1}^{N} m_{ai} \log m_{ai}$. The last term $-\zeta \sum_{i=1}^{M} \sum_{a=1}^{N} m_{ai}$ is used to penalize unmatched points. The minimization of $F(\{m_{ai}\},\mathbf{A},\mathbf{f})$ is done by imposing the two-sided correspondence constraints which transform into constraints on $\sum_{a=0}^{N} m_{ai} = 1, \ \forall i$ and $\sum_{i=0}^{N} m_{ai} = 1, \ \forall a$.

While this approach can give very successful results [8], it does need good initialize conditions, and can require many iterations to converge (see Section 5). Rangarajan et al. [30] have considered the alternative Bethe/Kikuchi

free energy approximation. This appears to yield good results, but the convergence rate is slow.

3. The discriminative model

The generative model described above is an attractive way to formulate the problem. But it has three computational disadvantages: (i) the initialization, (ii) the one-sided correspondence constraints required to perform the E-step, (iii) the convergence rate. These problems arise because the ambiguity of point features for matching.

To address these concerns, we develop a complementary discriminative model. This model makes use of shape features which exploit the local and global context of the shape. These shape features are motivated by shape contexts [4] and are far less ambiguous for matching than the points used in the generative model. We therefore call them *informative features*. The shape features for the continuous-contour representation exploit knowledge of the arc-length, and so will be more informative than those for the sparse-point representation. As will be described in Section 4, the shape features enable us to get good initialization and to obtain a practical E-step which gives rapid convergence (in combination with the M-step based on the generative model).

To formally specify the discriminative model, we define shape features on the target by $\phi(\mathbf{x}_i)$, and on the source by $\phi(\mathbf{A}\mathbf{y}_a + \mathbf{f}(\mathbf{y}_a))$. This implies that the features on the source shape depend on the current estimates of the geometric transformation $(\mathbf{A},\mathbf{f})$, which is initialized to be the identity. Intuitively, we are warping the source onto the target. We define a similarity measure between the shape features to be $q(.,.)$. The forms of ϕ and $q(.,.)$ will be defined in Sections 3.1 and 3.2 and will differ for the continuous-contour and sparse-point representations.

The discriminative model is given by $P_D(V,\mathbf{A},\mathbf{f}|X,Y) = P_D(V|X,Y,\mathbf{A},\mathbf{f})P(\mathbf{A})P(\mathbf{f})$, where the priors $P(\mathbf{A}),P(\mathbf{f})$ are defined as in Section 2.3, and:

$$-\log P_D(V,|X,Y,\mathbf{A},\mathbf{f}) = \sum_{a=1}^{M} q(\phi(\mathbf{x}_{v(a)}),\phi(\mathbf{A}\mathbf{y}_a + f(\mathbf{y}_a))) + \log Z_D. \quad (18)$$

This equation expresses the correspondence variables in terms of the one-sided variables $v(a)$. This has disadvantages, as described in Section 2.5. But these disadvantages are far less severe for the discriminative model because the shape features are far less ambiguous for matching than the point features. Note that this does not include a term allowing for unmatched points (analogous to the λ term in Eq. (14)). We experimented with such a term, but it did not make any significant difference.

The free energy for the discriminative model is defined to be:

$$F_D(Q,\mathbf{A},\mathbf{f}) = -\sum_V Q(V)\log P_D(V|X,Y,\mathbf{A},\mathbf{f}) - \log P(\mathbf{A}) - \log P(\mathbf{f}) + \sum_V Q(V)\log Q(V). \quad (19)$$

As before, the use of one-sided constraints means that we can factorize $Q(V)$ and compute it as:

$$Q(V_{ai}=1) = \frac{q(\phi(\mathbf{x}_i),\phi(\mathbf{A}\mathbf{y}_a+\mathbf{f}(\mathbf{y}_a)))}{\sum_{j=1}^{M} q(\phi(\mathbf{x}_i),\phi(\mathbf{A}\mathbf{y}+\mathbf{f}(\mathbf{y}_a)))}. \quad (20)$$

This gives the discriminative free energy:

$$F_D(Q,\mathbf{A},\mathbf{f}) = E_A(\mathbf{A}) + E_f(\mathbf{f}) + \log \hat{Z}_D - \sum_{a=1}^{N}\log\sum_{i=1}^{M} e^{-q(\phi(\mathbf{x}_i),\phi(\mathbf{A}\mathbf{y}_a+\mathbf{f}(\mathbf{y}_a))}. \quad (21)$$

We can also use Eq. (20) to calculate the E-step by setting the left-hand-side to be the distribution at time t with the right-hand-side evaluated at time $t-1$.

The M-step is more difficult since it requires differentiating the features ϕ and the matching term $q(.,.)$ with respect to $(\mathbf{A},\mathbf{f})$. The M-step will not be used in our complete algorithm.

This discriminative model has similarities to the shape features model [4]. The main difference is that the shape context model is not formulated in a probabilistic framework. Instead it requires that the source and target shapes have the same number of points and requires that each point has a unique match. This can be problematic when there are unequal numbers of points and can cause false matching. Our method is like softmax [8,42] and allows for uncertainty.

3.1. Shape features for the continuous-contour representation

The local and global features for the continuous-contour representation are illustrated in Fig. 2.

The **local features** at a point $\mathbf{x}(s_i)$ with tangent angle ψ_i are defined as follows. Choose six points on the curve by

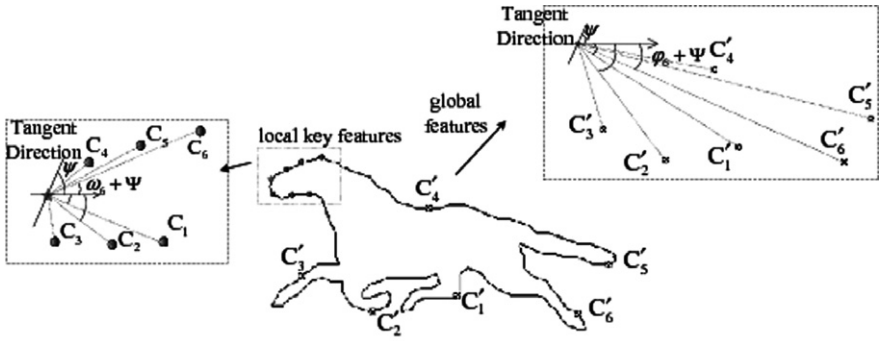
$$(\mathbf{x}(s_i-3\mathrm{d}s),\mathbf{x}(s_i-2\mathrm{d}s),\mathbf{x}(s_i-\mathrm{d}s),\mathbf{x}(s_i+\mathrm{d}s),\mathbf{x}(s_i+2\mathrm{d}s),\mathbf{x}(s_i+3\mathrm{d}s)), \quad (22)$$

where $\mathrm{d}s$ is a small constant ($\mathrm{d}s = 0.01$ in our experiments). The angles of these positions with respect to the point $\mathbf{x}_i$ are $(\psi_i+\omega_j, j=1,\ldots,6)$. The *local features* are $h_l(\mathbf{x}_i) = (\omega_j, j=1,\ldots,6)$.

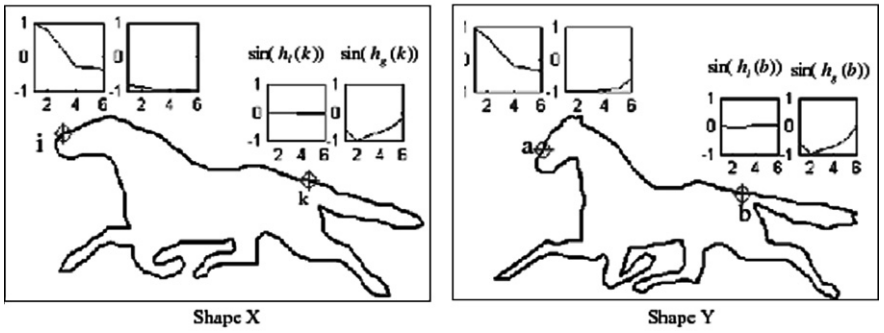
The **global features** are selected in a similar way. We choose six points near $\mathbf{x}(s_i)$, with tangent angle ψ_i, to be

$$(\mathbf{x}(s_i-3\Delta s),\mathbf{x}(s_i-2\Delta s),\mathbf{x}(s_i-\Delta s),\mathbf{x}(s_i+\Delta s),\mathbf{x}(s_i+2\Delta s),\mathbf{x}(s_i+3\Delta s)), \quad (23)$$

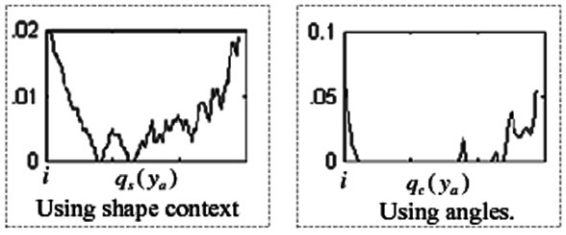
where Δs is a large constant ($\Delta s = 0.1$ in our experiments), with angles $\psi_i+\varphi_j, j=1,\ldots,6$. *The global features* are $h_g(\mathbf{x}_i) = (\varphi_j, j=1,\ldots,6)$. Observe that the features $\phi = (h_l,h_g)$ are invariant to rotations in the image plane.



(a) Local and global features for connected points.



(b) Similarity (affinity) measures on the features.



(c) Similarity (affinity) measures on the features.

Fig. 2. The features and the similarity measure. (a) Illustrates how the local and global features are measured for the complete-contour representation. The left window shows the angles computed at the local neighboring points which are used to compute the local shape features. The right window shows the angles computed in larger neightborhoods to compute the global features. (b) Displays the featuers of two points in shape X and Y. As angles have periodic property, we show the a angle vectors by their sinusoid values. As we can see, the corresponding points i on X and a on Y, and$k$on X and b on Y have similar features. The left and right figure in (c) plot the similarities between point a on Y with respect to all points in X using the shape context (on sparse point) feature and angle features (on complete-contour) respectively. As we can see similarities by features defined in this paper for connected points have lower entropy than these given by shape contexts. They are less ambiguous and more informative.

These shape features are informative for matching. Fig. 2b plots sinusoids ($\sin(h_l)$,$\sin(h_g)$) for two points on X and two points on Y. Observe the similarity between these features on the corresponding points.

The similarity measure between the two points is defined to be:

$$q_c(\phi(\mathbf{x}_i),\phi(\mathbf{y}_a)) = 1 - c_1\left\{\sum_{j=1}^{6} D_{angle}(\omega_j(\mathbf{x}_i) - \omega_j(\mathbf{y}_a)) + \sum_{j=1}^{6} D_{\text{angle}}(\varphi_j(\mathbf{x}_i) - \varphi_j(\mathbf{y}_a))\right\}, \quad (24)$$

where $D_{\text{angle}}(\omega_j(\mathbf{x}_i) - \omega_j(\mathbf{y}_a))$ is the minimal angle from $\omega_j(\mathbf{x}_i)$ to $\omega_j(\mathbf{y}_a)$ (i.e. modulo 2π), and c_1 is a normalization constant. The right panel in Fig. 2c plots the vector $q_c(\mathbf{y}) = [q_c(\phi(\mathbf{x}_i),\phi(\mathbf{y})), i = 1..M]$ as a function of i for points $\mathbf{y}_a$ and $\mathbf{y}_b$ on Y, respectively.

3.2. Shape features for the sparse-point representation

In this case, we also use local and global features. To obtain the local feature for point $\mathbf{x}_i$, we draw a circle with radius r (in our experiments $r = 9.0 \times \bar{r}$, where $\bar{r}$ is the average distance between points) and collect all the points that fall within the circle. The angles of these points relative to $\mathbf{x}_i$ are computed. The histogram of these angles is then used as the local feature, H_l. Observe that this representation is not invariant to rotation.

The global feature for the sparse points is computed by shape contexts [4] to give a histogram H_g (60 histogram

bins are used in the experiments). The local and global features are represented together as $\phi = (H_l, H_g)$.

The feature similarity between two points $\mathbf{x}_i$ and $\mathbf{y}_a$ is measured by the χ^2 distance on the histograms:

$$q_s(\phi(\mathbf{x}_i), \phi(\mathbf{y}_a)) = 1 - c_2\{\chi^2(H_l(\mathbf{x}_i), H_l(\mathbf{y}_a)) + \chi^2(H_g(\mathbf{x}_i), H_g(\mathbf{y}_a))\}. \quad (25)$$

To show the effectiveness of these features, the left panel of Fig. 2c plots the vector

$$q_s(\mathbf{y}_a) = [q_s(\phi(\mathbf{x}_i), \phi(\mathbf{y}_a)), i = 1 \ldots M] \quad (26)$$

as a function of i for a point $\mathbf{y}_a$ on Y.

The features for the sparse-point representation are generally less informative than those for the continuous-contour representation. For example, the shape context features [4] and histogram similarity measures tend to have high entropy, see Fig. 2c.

4. The full algorithm

Our complete algorithm combines aspects of the generative and the discriminative (data-driven) models. We first use the discriminative model to help initialize the algorithm, see Section 4.1. Then we use the M-step from the generative model and the E-step from the discriminative model.

Both steps are guaranteed to decrease the free energy of the corresponding model, but there is no guarantee that the M-step will decrease the free energy for the discriminative model—or that the E-step will decrease the free energy for the generative model. Nevertheless out computer simulations show that they always do.

4.1. Initialization

We use the discriminative model for initialization. In this paper, we only initialize with respect to θ since the rest of the algorithm was fairly insensitive to other global parameters such as scaling.

In theory, the initialization should be done by marginalizing over the irrelevant variables. To estimate the full affine transformation $\mathbf{A}$ we should compute $P(\mathbf{A}|X,Y)$, and to estimate the angle θ we should compute $P(\theta|X,Y)$. But these computations are impractical for either the generative or discriminative models.

Instead we use an approximation $P_{In}(\theta|X,Y)$ motivated by the discriminative model,

$$P_{In}(\theta|X,Y) = \sum_{i=1}^{M}\sum_{a=1}^{N} q(\phi(\mathbf{x}_i), \phi(\mathbf{y}_a))\delta(\theta - \theta(a,i,X,Y)), \quad (27)$$

where $q(.,.) = q_c(.,.)$ or $q_s(.,.)$ depending on whether we use the continuous-contour or sparse-point representation. Here, $\theta(a,i,X,Y)$ is the estimated angle if we match point i on X to point a on Y.

We evaluate the distribution $P_{In}(\theta|X,Y)$ using mean-shift clustering to obtain several modes. The peaks of this distribution gives estimates for the initial values of θ_{initial}. In rare cases, the distribution will have several peaks. For example, suppose we are matching the equilateral triangles shown in Fig. 3. In this case $P_{\text{In}}(\theta|X,Y)$ will have three peaks, and we will have to consider initializations based on each peak.

To initialize the continuous-contour representation, we can exploit the rotational invariance of the features. This means that we only need to evaluate them for a small set of angles. But for the sparse-point representation, the features are not invariant to rotation.

In rare cases, we will require to sum over several modes. For example, three modes $(\theta^*, \theta_1^*, \theta_2^*)$ are required when matching two equal lateral triangles, see Fig. 3.

4.2. The E step: approximating $Q(V)$

The features for the continuous-contour representation are altered after the initialization by incorporating the local tangent angle ψ_i at each point. This makes these features rotation dependent.

We also augment the similarity measure by including the scaled relative position of the point to the center of the shape. For a point $\mathbf{x}_i$ on the target shape, this gives $\mathbf{x}_i - \bar{\mathbf{x}}$ where $\bar{\mathbf{x}} = \frac{1}{M}\sum_i \mathbf{x}_i$. We compute a similar measure $\mathbf{y}_a - \bar{\mathbf{y}}$, with $\bar{\mathbf{y}} = \frac{1}{N}\sum_a \mathbf{y}_a$. The full similarity measure is defined to be:

$$q_c'(\phi(\mathbf{x}_i), \phi(\mathbf{y}_a)) = 1 - c_1\sum_{j=1}^{6}[D_{\text{angle}}(\omega_j(\mathbf{x}_i) - \omega_j(\mathbf{y}_a)) + D_{\text{angle}}(\varphi_j(\mathbf{x}_i) - \varphi_j(\mathbf{y}_a)))] - c_2\|(\mathbf{x}_i - \bar{\mathbf{x}}) - (\mathbf{y}_a - \bar{\mathbf{y}})\|^2. \quad (28)$$

where $c_1 = 0.7$ and $c_2 = 0.8$ in our implementation.

This gives a factorized probability model:

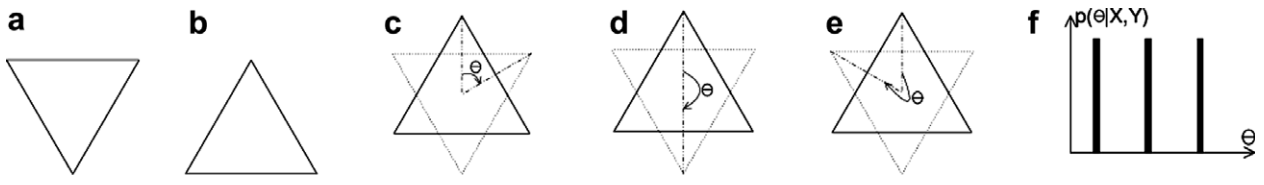


Fig. 3. The distribution it p($\theta|X,Y$), shown in (f), has three modes for a target shape X, shown in (a), and a source shape Y, shown in (b). The three possible values for θ are shown in (c), (d), and (e).

$$P_D(V|X,Y,\mathbf{A},\mathbf{f}) = \prod_a P_D(v_a|\mathbf{y}_a,X,\mathbf{A},\mathbf{f}). \quad (29)$$

where

$$P_D(v_a = i|\mathbf{y}_a,X,\mathbf{A},\mathbf{f}) = \frac{q'_c(\phi(\mathbf{x}_i),\phi(\mathbf{A}\mathbf{y}_a+\mathbf{f}(\mathbf{y}_a)))}{\sum_{j=1}^{M} q'_c(\phi(\mathbf{x}_j),\phi(\mathbf{A}\mathbf{y}_a+\mathbf{f}(\mathbf{y}_a)))}. \quad (30)$$

4.3. The M step: estimating $\boldsymbol{A}$ and $\boldsymbol{f}$

The M-step corresponds to minimizing the following energy function with respect to $(\mathbf{A},\mathbf{f})$:

$$E_M(\mathbf{A},\mathbf{f}) = \sum_i \sum_a P_D(i,a)\|\mathbf{x}_i - \mathbf{A}\mathbf{y}_a - \mathbf{f}(\mathbf{y}_a)\|^2 + E_A(\mathbf{A}) + E_f(\mathbf{f}), \quad (31)$$

where $P_D(i,a) = P_D(v_a = i|\mathbf{y}_a,X,A,\mathbf{f})$.

Recall that we write $\mathbf{f}(\mathbf{y}_a) = \mathbf{t} + \sum_{b=1}^{M}\mathbf{w}_b U(\mathbf{y}_a - \mathbf{y}_b)$. We re-express the energy as:

$$\begin{aligned} E_M(\mathbf{A},\{\mathbf{w}_b\},\mathbf{t}) = &\sum_i \sum_a P_D(i,a)\|\mathbf{x}_i - \mathbf{A}\mathbf{y}_a - \mathbf{t} \\ &- \sum_{b=1}^{M}\mathbf{w}_b U(\mathbf{y}_a - \mathbf{y}_b)\|^2 + E_A(\mathbf{A}) \\ &+ \lambda \sum_{a,b=1}^{N} \mathbf{w}_b^T U(\mathbf{y}_b - \mathbf{y}_a)\mathbf{w}_a. \end{aligned} \quad (32)$$

We take the derivatives of E_M with respect to $\mathbf{A},\{\mathbf{w}_b\},\mathbf{t}$ and set them to zero:

$$\begin{aligned} &\sum_{i=1}^{M}\sum_{a=1}^{N} P_D(i,a)\{\mathbf{A}\mathbf{y}_a + \mathbf{t} + \sum_{b=1}^{N}\mathbf{w}_b U(\mathbf{y}_a - \mathbf{y}_b)\} = \sum_{i=1}^{M}\sum_{a=1}^{N} P_D(i,a)\mathbf{x}_i, \\ &\lambda\sum_{c=1}^{M} U(\mathbf{y}_c - \mathbf{y}_b)\mathbf{w}_c + \sum_{a=1}^{N}\sum_{i=1}^{M} P_D(i,a)U(\mathbf{y}_a - \mathbf{y}_b)\{\mathbf{A}\mathbf{y}_a + \mathbf{t} \\ &+ \sum_{c=1}^{M} U(\mathbf{y}_a - \mathbf{y}_c)\mathbf{w}_c\} = \sum_i \sum_a P_D(i,a)U(\mathbf{y}_a - \mathbf{y}_b)\{\mathbf{x}_i - \mathbf{A}y_a\}, \end{aligned} \quad (33)$$

$$\begin{aligned} &\sum_{i=1}^{M}\sum_{a=1}^{N}\sum_{\rho=1}^{2} A_{\mu\rho}y_{a,\rho}y_{a,v} + (1/2)\frac{\partial E_A}{\partial A_{\mu v}} + \sum_{i=1}^{N}\sum_{a=1}^{M}\sum_{b=1}^{M} P_D(i,a) \\ &\times w_{b,\mu}U(\mathbf{y}_a - \mathbf{y}_b) = \sum_{i=1}^{M}\sum_{a=1}^{N} P_D(i,a)x_{i,\mu}. \end{aligned} \quad (34)$$

Here μ,v,ρ are the spatial indices, the partial derivative of E_A is evaluated at the previous state.

Eq. (34) gives a set of linear equations in $\mathbf{A},\{\mathbf{w}_b\},\mathbf{t}$, which can be solved by complicated variant of the techniques described in [4].

4.4. Summary of the algorithm

The algorithm proceeds as follows:

(1) Given a target shape X and a source shape Y, it computes their shape features and estimates the rotation angle(s) θ_0.

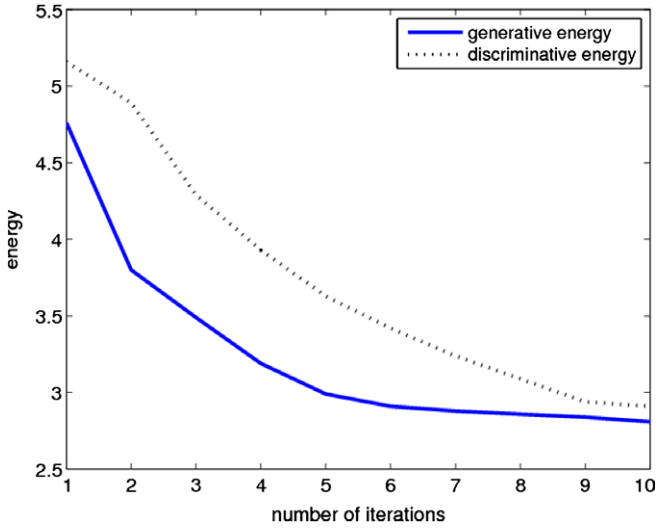
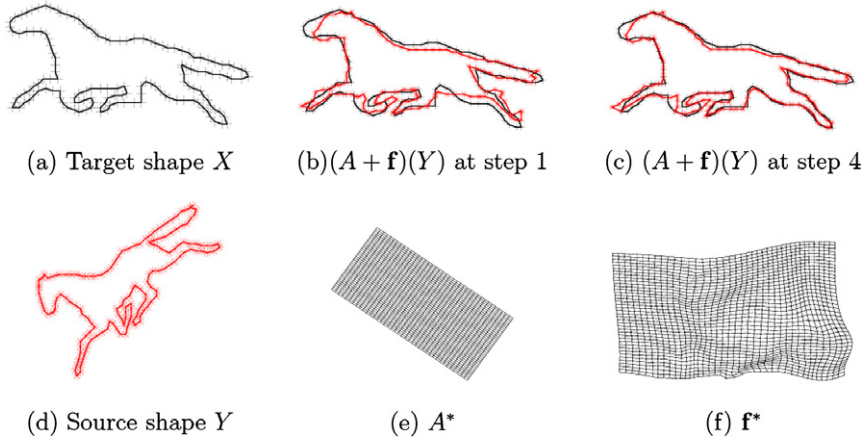


Fig. 5. Illustration of the decrease of generative free energy and discriminative free energy by our algorithm. We see that both the free energies consistently decrease w.r.t time.



(a) Target shape X (b) $(A+\mathbf{f})(Y)$ at step 1 (c) $(A+\mathbf{f})(Y)$ at step 4

(d) Source shape Y (e) A^* (f) $\mathbf{f}^*$

Fig. 4. The dynamics of the algorithm for the horse example.

(2) For each rotation angle θ_0, rotate the shape Y by θ_0 to Y_{θ_0}. Initialize the remaining transformations to be the identity.
(3) Estimate the features for the rotated shape Y_{θ_0}. Use these to initialize $Q(V)$ using the E-step of the discriminative model.
(4) Calculate $(\mathbf{A},\mathbf{f})$ by performing the M-step for the generative model (using a quadratic approximation).
(5) Transform Y by $(\mathbf{A},\mathbf{f})$, then repeat the last two stages for M iterations.
(6) Compute the similarity measure $P(X,\mathbf{A}^*,\mathbf{f}^*|Y)$ according to Eq. (6) and keep the best $(\mathbf{A}^*,\mathbf{f}^*)$.

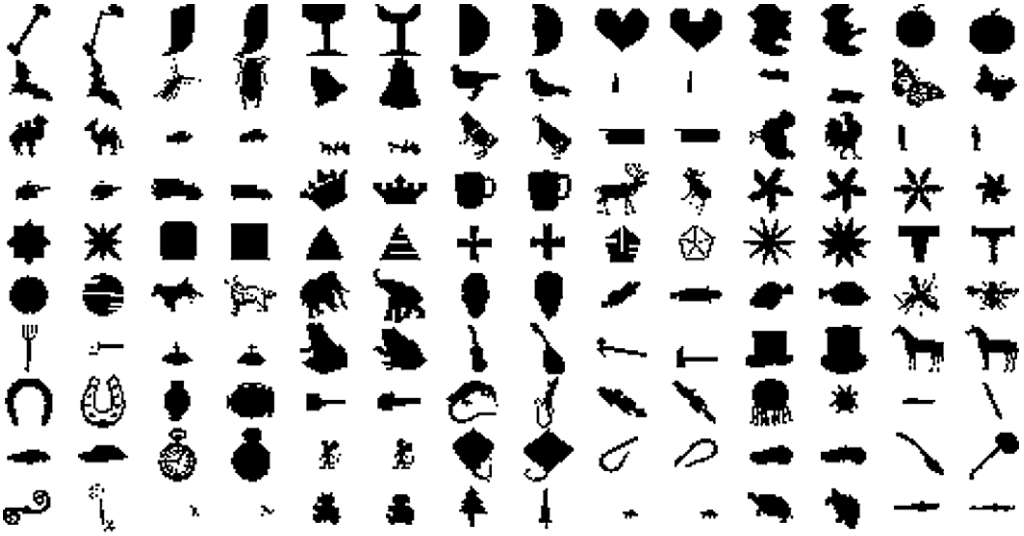
In practice, we found that $M = 4$ iterations was typically a sufficient number of iterations because of the effectiveness of the features (the shape context algorithm [4] uses $M = 3$). We needed $M = 10$ iterations for the experiments in Section 5.3. The algorithm runs at 0.2 s for matching X and Y of around 100 points on PC with 2.0 Ghz.

Fig. 4 gives an example of the algorithm where the source shape Y in (d) is matched with the target shape X. Fig. 4e and f show the estimated transformation $\mathbf{A}^*$ and $\mathbf{f}^*$.

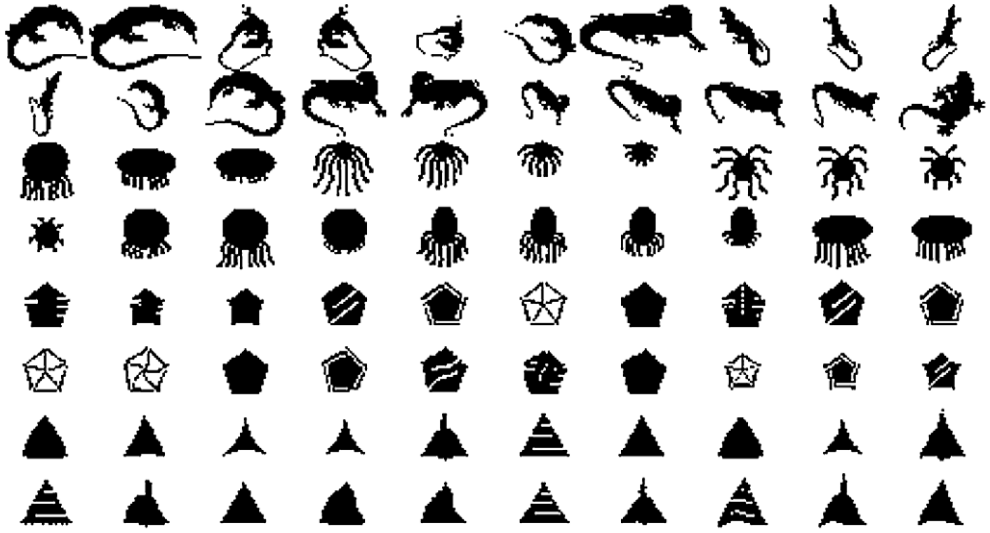
Fig. 5 shows the generative and discriminative free energies plotted for the example. Observe that both decrease monotonically.

5. Experiments

We tested our algorithm on standard shape matching datasets and for grey level image tasks such as registration, recognition, and foreground/background segmentation. We use



(a) Some typical images in MPEG7 CE-Shape-1



(b) The four types with the lowest rates

Fig. 6. The MPEG7 CE-Shape-1.

Table 1
The retrieval rates of different algorithms for the MPEG7 CE-Shape-1

Algorithm	CSS	Visual parts	Shape contexts	Curve edit distance	Our method
Recognition rate	75.44%	76.45%	76.51% [4]	78.17% [32]	80.03%

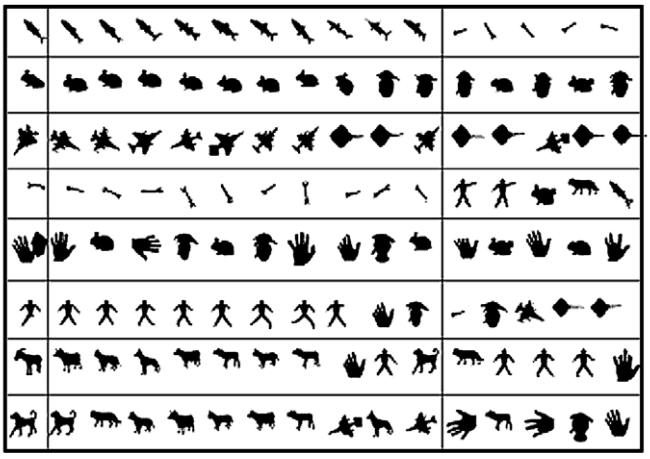
The results by the other algorighms are taken from Sebastian et al. [32].

the continuous-contour representation for examples 5.1, 5.2, and 5.4 and the sparse-point representation for the remainder.

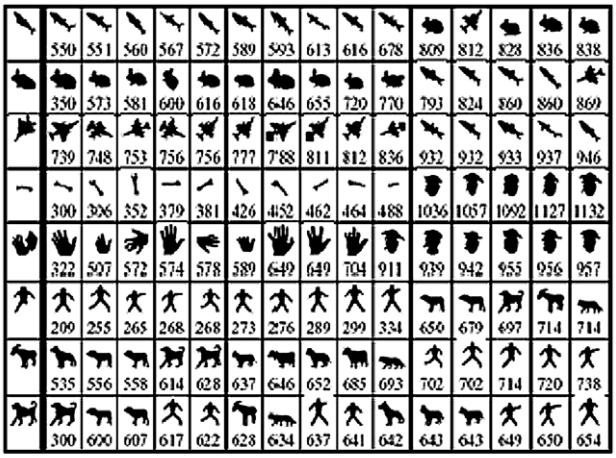
5.1. MPEG7 shape database

We first tested our algorithm on the MPEG7 CE-Shape-1 [18] database. This database contains 70 types of objects each of which has 20 different silhouette images, giving a total of 1400 silhouettes. Since the input images are binarized, we can extract the contours and use the continuous-contour representation. Fig. 6a displays two images for each object type. The task is to do retrieval and the recognition rate is measured by the Bull's eye criterion [18]. A single shape is presented as a query and the top 40 matches are obtained.

(a) The 99 silhouette images of the Kimia data set.



(b) Some matching results by our method



(c) Some matching results by Shock Edit

Fig. 7. The Kimin data set fo 99 shapes and some matching results.

This is repeated for every shape and the number of correct matches (out of a maximum total of 20) are obtained.

The recognition rates for different algorithms are given in Table 1. On this dataset our algorithm outperforms the alternatives. The speed is in the same range as those of shape contexts [4] and curve edit distance [32].

5.2. The Kimia data set

We then tested the identical algorithm (i.e. continuous-contour representation and with the same algorithm parameters) on the Kimia data set of 99 shapes [33]. These are shown in Fig. 7a. For each shape, the 10 best matches are picked (since there are 10 other images in the same category). Table 2 shows the numbers of correct matches. Our method performs similarly to Shock Edit [33] for the top 7 matches, but is worse for the remainder. Shape context performs worse than either algorithm on this task. Fig. 7b shows the 15 top matches for some shapes.

The reason that our algorithm does worse than Shock Edit on the last three examples is because the geometric deformations involve the presence or absence of parts, see Fig. 7. Our model does not take these types of geometric transformation into account. Dealing with them requires a more sophisticated shape representation.

5.3. Chui and Rangarajan data set

We tested our algorithm on data supplied by Chui and Rangarajan [8]. This data includes noise points in the target shape. We use the sparse-point representation since we do not have a closed contour for each shape.

The algorithm runs for 10 steps for this dataset and results are given in Fig. 8. The quality of our results are similar to those reported in [8]. But our algorithm runs an estimated 20 times fewer iteration steps.

5.4. Text image matching

The algorithm was also tested on real images of text in which binarization was performed by the method described

Table 2
The number of matched shapes by the different algorithms

Algorithm	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6	Top 7	Top 8	Top 9	Top 10
Shock edit	99	99	99	98	98	97	96	95	93	82
Our method	99	97	99	98	96	96	94	83	75	48
Shape contexts	97	91	88	85	84	77	75	66	56	37

The results by the other algorithms are due to Sebastian et al. [33].

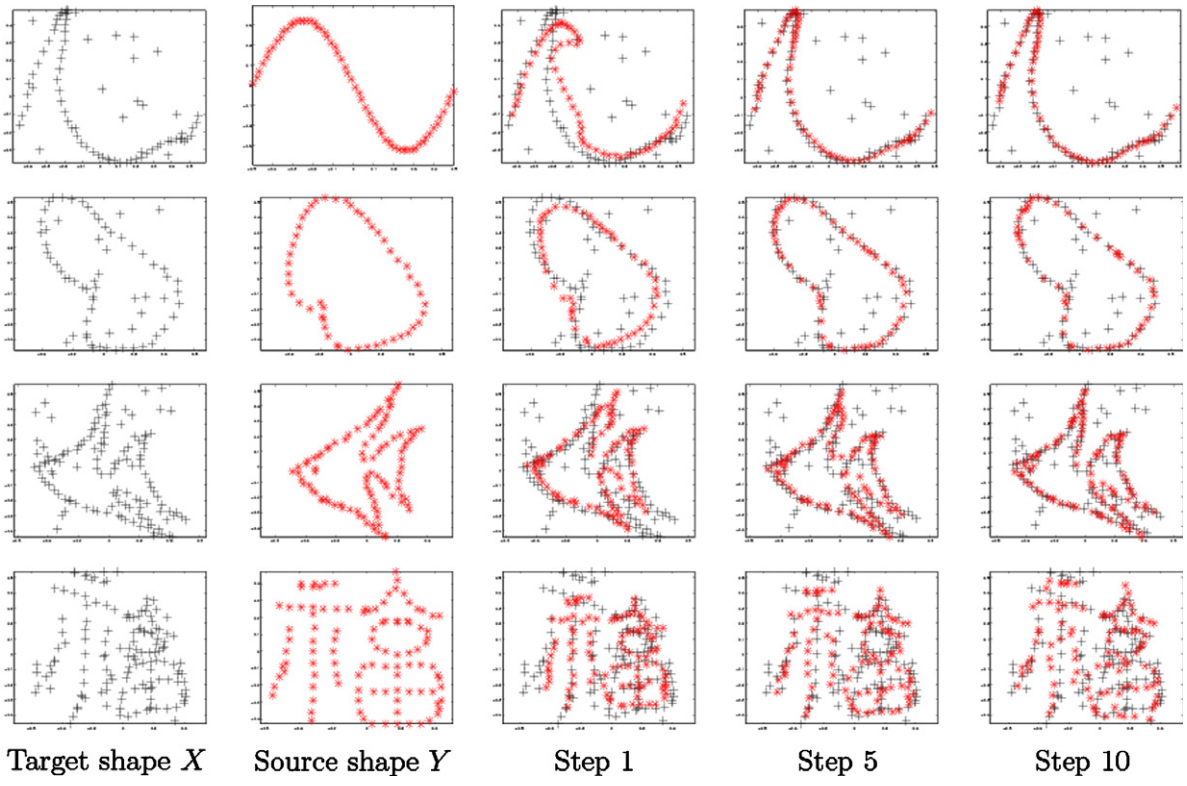


Fig. 8. Some results on the Chui and Rangarajan data set.

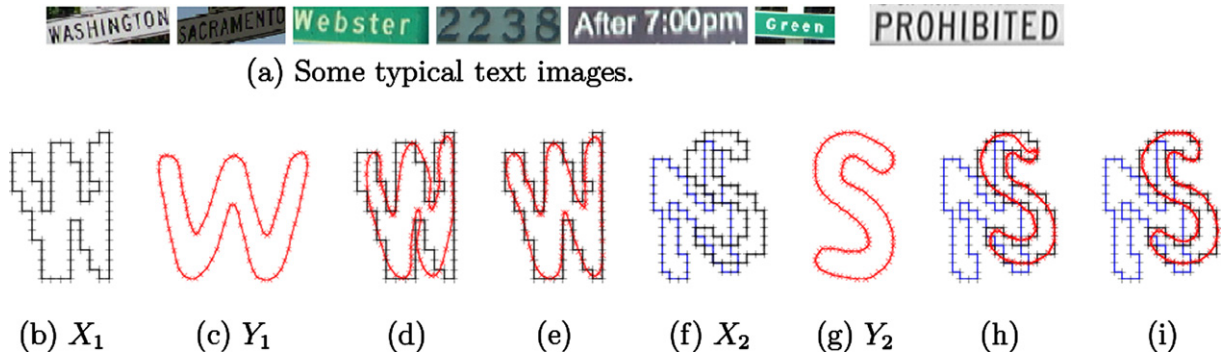


Fig. 9. Results on some text images, (c) and (i) display the matching. We purposely superimpose two shapes extracted from different images together and make a new input. Our algorithm is robust in this case.

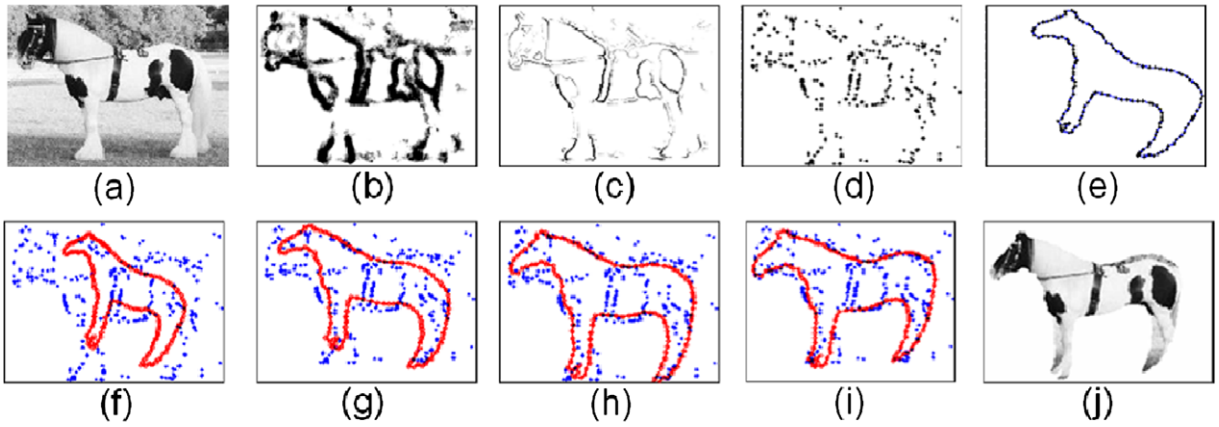


Fig. 10. Foreground/background segmentation: (a) input image; (b) the probability map BEL; (c) the product of BEL map and the canny edge map; (d) sampled points according to the probability (c); (e) the template; (f) (i) the result after the first. The third. The 7th and 10th iteration; (j) the segmented foreground according to the matching result.

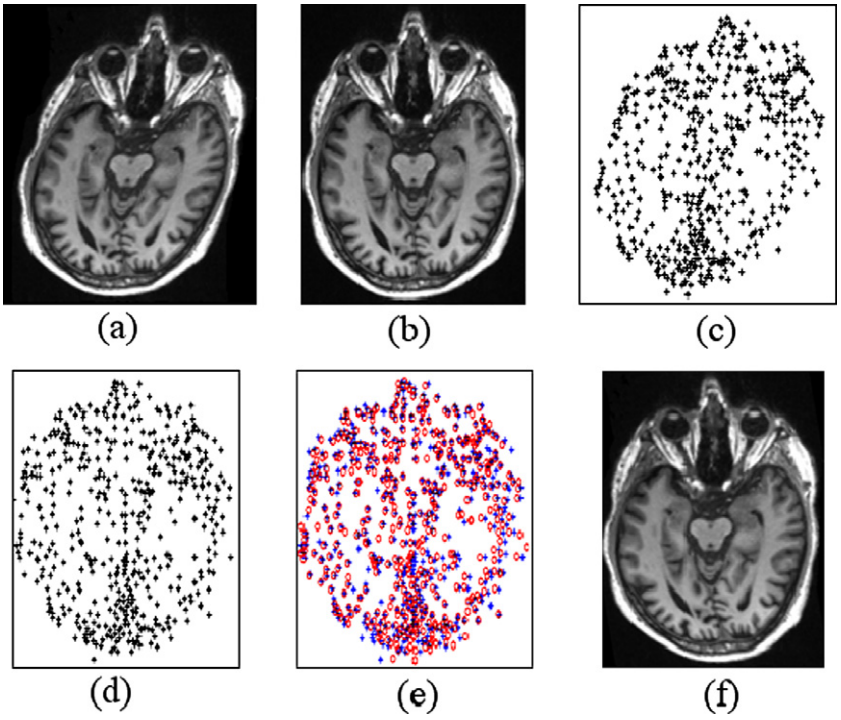


Fig. 11. Medical image registration: (a) the unregistered image: (b) the telplate image: (c) (d) the SIFT key points for the unregistered and template images; (c) the matching result; (f) the registered image.

in [7], followed by boundary extraction. This enabled us to perform shape matching using the continuous-contour representation. Several examples are shown in Fig. 9. Similar results can be obtained by matching the model to edges in the image.

5.5. Foreground/background segmentation

Our next example is a foreground/background segmentation task. The input is a grey level image with a horse in the foreground. In this case, binarization and edge detection techniques are unable to estimate the boundary of the horse. Instead we use a novel boundary detector named Boosted Edge Learning (BEL) [10] which is learnt from training data and makes use of local image context. BEL is applied to all points on the Canny edge map of the image. The output of BEL is a probability map which gives the probability that each pixel is on the edge of the horse. We sample from this probability map to obtain 300 points which we use as the sparse-point representation of the target shape.

The source shape is represented by 300 points randomly sampled from the silhouette of a horse. Fig. 10 shows an example on horse data [6].

5.6. Application to medical image registration

Finally we applied our algorithm to the task of registering medical images. This is an important practical task

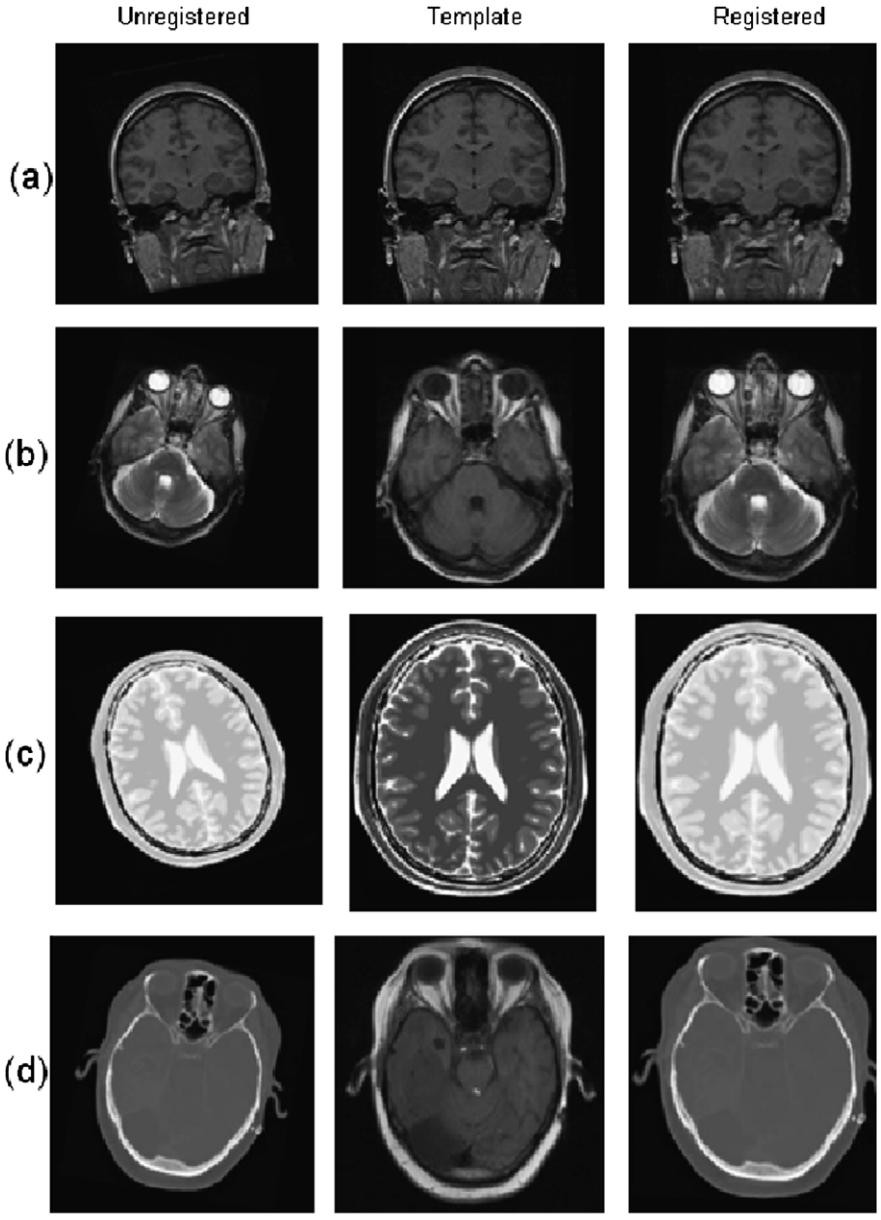


Fig. 12. Examples of registration results; (a) MRI registration; (b) registration and unregistered T2 image to T1 image; (c) registration an unregistered CT image to an MR image; (d) registration and unregistered PD image to T2 image.

which has been addressed by a variety of methods [23] including mutual information [22,29,40].

In this example, we used SIFT [20] to detect interest-points which we use as sparse-point representations for the target and source shapes. Then we applied our shape matching algorithm. The SIFT features are invariant to rotation and scaling, so we do not need to recalculate them as the algorithm proceeds.

Fig. 11 shows the registration process using our shape matching method. Fig. 12 shows more results of registration between different modalities of images, such as Computed Tomography (CT) versus Magnetic Resonance Imaging (MRI), Proton Density (PD) versus MRI, and we see that the registered results are very good.

Previous authors have applied SIFT to medical image registration [25,38]. But these papers used SIFT features and the shape context model [4] and so did not take uncertainty into account.

6. Conclusion

In this paper, we presented an efficient algorithm for shape matching, recognition, and registration. We defined a generative and discriminative models and discussed their relationships to softassign [8] and shape contexts [4]. We used an EM algorithm for inference by exploiting properties of both models. A key element is the use of informative shape features to guide the algorithm to rapid and correct solutions. We illustrated our approach on datasets of binary and real images, and gave comparison to other methods. Our algorithm runs at speeds which are either comparable to other algorithms or considerably faster.

Our work is currently limited by the types of representations we used and the transformations we allow. For example, it would give poor results for shape composed of parts that can deform independently (e.g., human figures). For such objects, we would need representations based on symmetry axes such as skeletons [32] and parts [44].

Acknowledgments

This project was done when Zhuowen Tu was a post-doc at UCLA supported by NIH (NEI) Grant 1R01EY013875. We thank Dr. David Lowe for sharing the SIFT source code online. We thank comments from anonymous reviewers for improving the clarity of the paper.

References

[1] S. Abbasi and F. Mokhtarian, Robustness of Shape Similarity Retrieval under Affine Transformation, in: Proceedings of Challenge of Image Retrieval, Newcastle upon Tyne, UK. 1999.

[4] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (24) (2002) 509–522.

[5] F.L. Bookstein, Principal Warps: Thin-plate splines and the decomposition of deformations, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (6) (1989) 567–585.

[6] E. Borenstein, E. Sharon, and S. Ullman, Combining top-down and bottom-up segmentation, in: Proceedings of IEEE Workshop on Perceptual Organization in Computer Vision, vol. 4, pp. 46–53, June 2004.

[7] X. Chen and A.L. Yuille. AdaBoost Learning for Detecting and Reading Text in City Scenes, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 366–373, Washington, DC, June 2004.

[8] H. Chui, A. Rangarajan, A new point matching algorithm for non-rigid registration, Computer Vision and Image Understanding 89 (2003) 114–141.

[9] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum-likelihood from incomplete data via the EM algorithm, Journal of Royal Statistics Society Series B 39 (1977).

[10] P. Dollár, Z. Tu, and S. Belongie, Supervised learning of edges and object boundaries, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1964-1971, New York, June 2006.

[11] D. Geiger, A.L. Yuille, A common framework for image segmentation, International Journal of Computer Vision 6 (3) (1991) 227–243.

[12] U. Grenander, General Pattern Theory: A Mathematical Study of Regular Structures, Oxford, 1994.

[14] M.K. Hu, Visual pattern recognition by moment invariants, IEEE Transactions Information Theory 8 (1962) 179–187.

[15] Y. Keselman, A. Shokoufandeh, M. Demirci, and S. Dickinson, Many-to-Many Graph Matching via Metric Embedding, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Madison, WI, pp. 850–857, 2003.

[18] L.J. Latechi, R. Lakamper, U. Eckhardt, Shape descriptors for non-rigid shapes with a single closed contour, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition 1 (2000) 424–429.

[19] S. Loncaric, A survey of shape analysis techniques, Pattern Recognition 31 (8) (1998) 983–1001.

[20] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[21] B. Luo, E.R. Hancock, A unified framework for alignment and correspondence, Computer Vision and Image Understanding 92 (1) (2003) 26–55.

[22] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, P. Suetens, Multimodality image registration by maximization of mutual information, IEEE Transactions on Medical Imaging 16 (2) (1997) 187–198.

[23] J.B. Maintz, M.A. Viergever, A survey of medical image registration, Medical Image Analysis 2 (1) (1998) 1–36.

[24] F. Mokhtarian, A.K. Mackworth, A Theory of Multiscale, Curvature-based shape representation for planar curves, IEEE Transactions on Pattern Analysis and Machine Intelligence 14 (8) (1992) 789–805.

[25] M. Moradi, P. Abolmaesoumi and P. Mousavi, Deformable Registration Using Scale Space Keypoints, Medical Imaging 2006: Image Processing, in: Proceedings of the SPIE, vol. 6144, pp. 791–798, 2006.

[26] R. Neal, G.E. Hinton, A view of the EM algorithm that justifies incremental, sparse, and Other Variants, in: Learning in Graphical Models, MIT Press, 1998, pp. 355–368.

[27] N. Paragios, M. Rousson, V. Ramesh, Matching distance functions: a shape-to-area variational approach for global-to-local registration, Proceedings of European Conference on Computer Vision 2 (2002) 775–788.

[28] M. Pelillo, K. Siddiqi, S.W. Zucker, Matching hierarchical structures using association graphs, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (11) (1999) 1105–1119.

[29] J.P. Pluim, J.B.A. Maintz, M.A. Viergever, Mutual-information-based registration of medical images: a survey, IEEE Transactions on Medical Imaging 22 (8) (2003) 986–1004.

[30] A. Rangarajan, J.M. Coughlan, A.L. Yuille, A bayesian network for relational shape matching, Proceedings of International Conference on Computer Vision 1 (2003) 671–678.

[32] T.B. Sebastian, P.N. Klein, B.B. Kimia, On aligning curves, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (1) (2003) 116–125.

[33] T.B. Sebastian, P.N. Klein, B.B. Kimia, Recognition of shapes by editing their shock graphs, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (5) (2004) 550–571.
[35] A. Thayananthan, B. Stenger, P.H.S. Torr, R. Cipolla, Shape context and chamfer matching in cluttered scenes, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition 1 (2003) 127–133.
[36] Z. Tu, X. Chen, A. Yuille, S.C. Zhu, Image parsing: unifying segmentation, detection and recognition, International Journal of Computer Vision 63 (2) (2005) 113–140.
[37] Z. Tu, A. Yuille, Shape matching and recognition—using generative models and informative features, Proceedings of European Conference on Computer Vision III (2004) 195–209.
[38] M. Urschler, J. Bauer, H. Ditt, and H. Bischof, SIFT and Shape Context for Feature-Based Nonlinear Registration of Thoracic CT Images, International ECCV Workshop on Computer Vision Approaches to Medical Image Analysis, pp. 73–84, 2006, Graz, Austria.
[39] R.C. Veltkamp and M. Hagedoorn, State of the Art in Shape Matching, Technical Report UU-CS-1999-27, Utrecht, 1999.
[40] W.M. Wells III, P. Viola, H. Atsumi, S. Nakajima, R. Kikinis, Multimodal volume registration by maximization of mutual information, Medical Image Analysis 1 (1) (1996) 35–52.
[41] A.L. Yuille, N.M. Grzywacz, A computational theory for the perception of coherent visual motion, Nature 333 (6168) (1988) 71–74.
[42] A.L. Yuille, Generalized deformable models, statistical physics, and matching problems, Neural Computation 2 (1) (1990) 1–24.
[43] C.T. Zahn, R.Z. Roskies, Fourier descriptors for plane closed curves, IEEE Transactions on Computers C-21 (3) (1972) 269–281.
[44] S.C. Zhu, A.L. Yuille, FORMS: a flexible object recognition and modeling system, International Journal of Computer Vision 20 (3) (1996) 187–212.