Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computer Vision and Image Understanding 99 (2005) 241–258
Computer Vision and Image Understanding
www.elsevier.com/locate/cviu

Object recognition with uncertain geometry and uncertain part detection

Thang V. Pham *, Arnold W.M. Smeulders

ISIS, Informatics Institute, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands

Received 14 July 2004; accepted 20 January 2005
Available online 10 March 2005

Abstract

This paper presents a method for object recognition once parts have been detected. The recognition task is formulated as a graph problem searching for the characteristic geographical arrangements of (possibly missing) parts. The objective function is Bayesian maximum a posteriori estimation, integrating the image likelihood as a posteriori probability of the part detectors. The variability in the arrangement of object parts is captured by a Gaussian distribution after translation normalization. By employing two special properties of a Gaussian distribution, we are able to deal with missing parts situation where the chosen origin is not detected. We use an A^* algorithm to find the optimal solution for the graph search problem. Experiments are performed on both synthetic and real data to demonstrate good results and fast performance of the recognition.

Keywords: A^* algorithm; Detection by parts; Detector performance; Graph search; Grouping; Maximum a posteriori estimation; Object recognition; Shape

1. Introduction

A major problem in computer vision is object recognition as the accidental circumstances of the scene and the recording dominate the appearance of the ob-

* Corresponding author. Fax: +31 20 525 7490.
E-mail address: vietp@science.uva.nl (T.V. Pham).

1077-3142/$ - see front matter
doi:10.1016/j.cviu.2005.01.006

ject. Apart from the variability of the object, the variability of the light source including its position and viewpoint, occlusion, noise, and clutter require careful consideration. Despite advances of systems recognizing specific objects such as [1], their use in real world applications is still limited, as these methods effectively require standardized recording circumstances. Recently, the problem of recognition of object class has received considerable attention because it is useful in a wide range of applications such as image retrieval, surveillance and robot navigation. Here, we consider objects with uncertain geometry and uncertain appearance. The uncertainty may come from individual object measurement error as well as from membership of a particular category such as pedestrians or human faces.

This paper investigates recognition using detectors of object parts. In this approach the object parts are first detected, then grouped to form objects according to an explicit global spatial relationship among parts [2]. This approach aims to overcome two problems of global description as of appearance-based methods [3,4]. First, a single template is not able to represent many objects without the inclusion of background where one wishes to detect objects and object parts without the accidental background. Experimental study in [5] has showed that the background affects adversely the performance of appearance-based recognition in a general context. Second, a global template allows only rather fixed structures.

The idea of recognition by parts is not new. Grimson [6] provides a complete account for methods that use geometrical constraints for describing the configuration. Nevertheless, the use of geometrical constraints among parts is not suited for describing the global variability of object configuration. Recently, developments in statistical shape theory have proved to be useful in modeling the global configuration [7–11] where the locations of parts (salient points) of an object are treated as a feature vector and subsequently statistical analysis can be carried out for a collection of vectors in this feature space. The work of Burl and Perona [2] is representative for the class of methods using these types of global deformable configuration to describe object structure. The method makes use of a shape space, obtained from original space by normalizing for translation, rotation and scale, in which the distribution can be expressed in a closed form, when the original space is Gaussian [12]. However, there is a difficulty in estimating the parameters of the distribution in their shape space. Another drawback of this method is that there is no guarantee that the search procedure will lead to an optimal solution.

We identify three issues in recognition by parts. The first issue is to determine the interesting object parts and their detectors from a collection of examples. The second issue is to integrate the part detection results as well as the variability of the configuration into a single model. Subsequently, the parameters of this model can be estimated from examples. The third issue is to solve the computational problem of finding the optimal solution according to the estimated model for object localization in a new input image.

We do not address the first issue in this paper. Automatic part selection is a parallel track of research where for example [13] shows some promising results. For

learning part detectors, one could make use of various techniques in the appearance-based approach. For the purpose of this paper, we use the standard template matching as a part detector.

We focus on the modeling and recognition problem given the part detectors. The difficulty lies in the fact that the scene is cluttered, possibly occluded, and part detectors are not reliable. Therefore, we will consider the issue of false and missed detection of object parts. In addition, unlike appearance-based approach where it is possible to conduct an exhaustive search in the space of all possible candidates, the hypothesis space in recognition by parts explodes very quickly.

We propose a Bayesian approach to integrate the distinctiveness of the object global configuration and the presence of its parts in a complex scene. The Bayesian formulation has been used for the object recognition problem in [14–18]. The advantage of Bayesian analysis is that prior knowledge can be incorporated into the evaluation procedure. Our work differs from previous work in the model we propose for the problem of recognition by parts. After normalized for translation, the variation in configuration of an object is modeled as multivariate Gaussian distribution. The performance of the part detectors is used to determine the relative weight of the parts. The performance parameters are the detection rate and the false alarm rate, which can be estimated from a set of part and non-part examples. Both the configuration and the weights are combined into a single Bayesian objective function to rank object hypotheses. This criterion leads to a difficult combinatoric problem because of the unconstrained covariance matrix of the Gaussian distribution. The structure of our probabilistic model is neither tree-based nor chain-based, which prevents the use of dynamic programming as it has been done in previous works [19,20]. Nevertheless, we will show that optimizing the new objective function can still be done efficiently with an A^* search algorithm [21], whereby an optimal solution is achieved with a novel, admissible heuristic.

For the optimization of the posterior, stochastic optimization algorithms such as simulated annealing and sampling have been successfully used by Refs. [22,23,16]. This class of algorithms is useful, but generally does not yet satisfy the computation time constraint [24]. A gradient descent method can be used to obtain a local solution [14]. For the specific probabilistic model in [25], the authors present a tree search procedure to optimize the posterior. This model is further studied in [24] in relation with the A^* search algorithm. Nevertheless, to use the A^* algorithm with a probabilistic model, one has to derive appropriate heuristics for that specific model. In this respect, our solution to the problem of optimizing the posterior is novel.

The organization of the paper is as follows. In Section 2 we present a Bayes formulation of the problem. In Section 3 we derive the objective function and discuss the estimation of its parameters. The subsequent section is devoted to the search method for our objective function. Section 5 presents our experimental results on both synthetic and real data. Finally, we discuss various issues for further investigation and conclude the paper in Section 6.

2. Bayes formulation of the problem

Consider an object model consisting of two components: the parts and their locations. For a set of p parts, p is fixed a priori, we consider a spatial configuration x of p locations in some configuration space $\mathcal{X}$. Thus, x is a point in a $2p$-dimensional space describing the locations of p labeled parts.

Let $\{d_\alpha | \alpha = 1, \ldots, p\}$ be a set of detectors corresponding to the set of parts. When d_α is applied to an image I in some image space $\mathcal{I}$, the response $d_\alpha(I)$ is a binary image, indicating any detected presence of part α. Let $d_\alpha^{\mathbf{v}}(I)$ denote the response at location $\mathbf{v} \in I$, a value of 1 for the presence of part and 0 otherwise. We assume that by local maximization of the response the detector provides one location of α in a small neighborhood. In other words, each hit in $d_\alpha(I)$ corresponds to one detected part α.

We choose to use the maximum a posteriori (MAP) estimate in deriving the objective function. Given an image $I \in \mathcal{I}$, one would like to find the "most probable" object. This can be formulated as

$$x^* = \arg\max_x P(x|I) = \arg\max_x P(I|x)P(x) \tag{1}$$

It is often convenient to consider the log form of Eq. (1)

$$x^* = \arg\max_x \{\log(P(I|x)) + \log(P(x))\}. \tag{2}$$

In other words, $\log(P(I|x)) + \log(P(x))$ is used as an objective function to evaluate a hypothesis x.

In this model, $P(I|x)$ is the likelihood measure that an image I is observed given the presence of an object at some specified location x. The estimation of this measure is related to statistics of natural images $P(I)$, where a set of filters is often used. In that case, $P(I)$ is characterized by the distribution (or histogram) of the filter responses [26,27]. The method described in [16] also estimates the image (observation) likelihood given an object by means of filter banks. These filter banks aim at capturing the distribution of responses of both background and foreground. In Section 3.1, we present our estimation method when only the performance statistics of the part detectors is available.

The prior density distribution $P(x)$ reflects our knowledge about the object. As an example, suppose x consists of two components $x = (x_S, x_T)$, denoting a translation-normalized shape x_S and a translation vector to position the object in the image x_T. Suppose that the two components are independent. Thus, one can write $P(x) = P(x_S)P(x_T)$. We model the prior knowledge about the object location via $P(x_T)$, which is useful for datasets where objects of interest have a high or low probability of presence in some region. In this paper, we assume a uniform distribution for $P(x_T)$, that is, no a priori knowledge about object location in an image. In Section 3.2, we present a translation invariant density distribution for the global configuration, which does not assume a Gaussian distribution for the original space as in [2]. This enables simple estimation of model parameters. The scale and rotation invariant properties are not dealt with. When the scale and rotation of the object

are involved, one needs to address those aspects not only for shape but also for part detectors. This presents a significant additional complexity.

Finally, we address the optimization problem of Eq. (1). This problem is difficult because of a large search space. In Section 4, we formulate the problem as a graph search problem. We then present an *admissible* heuristic required to use the A^* search algorithm to find the optimal solution.

In summary, we are concerned with the two terms in Eq. (2) and the optimization problem. In the context of this paper, we need to approximate the first term with the information extracted from the part detectors. The second term should be modeled in such away that captures the distinctiveness of the object global configuration. Finally, the optimization method should be efficient because a very large search space is expected.

3. Learning an object model

We present the likelihood and the prior terms of the model in Eq. (2). We then discuss the estimation of model parameters from training examples.

3.1. The likelihood

In this section, we estimate the likelihood $P(I|x)$ that the image I is observed, given an observation of object x. We use the performance statistics of part detectors for our estimation, and hence the likelihood becomes $P(\{d_\alpha(I)\}_{1\leqslant\alpha\leqslant p}|x)$, where p is the number of object parts and $d_\alpha(I)$ is the binary response image. For the purpose of this paper, assuming the part detectors are independent, we have

$$P(\{d_\alpha(I)\}_{1\leqslant\alpha\leqslant p}|x) = \prod_{1\leqslant\alpha\leqslant p} P_\alpha(d_\alpha(I)|\boldsymbol{x}_\alpha), \tag{3}$$

where $\boldsymbol{x}_\alpha$ is the vector of the true location of part α in x.

This assumption is not realistic because object parts might bear great similarity, e.g., the left eye and the right eye of human faces. Despite of this fact, the assumption is needed to reduce computational complexity.

For each part α and its detector d_α, let γ_α be the detection rate, that is $\gamma_\alpha = P(d^{\boldsymbol{v}}_\alpha(I) = 1|\boldsymbol{v} = \boldsymbol{x}_\alpha)$, where $\boldsymbol{v}$ is a two-dimensional vector denoting an image location. Thus, $P(d^{\boldsymbol{v}}_\alpha(I) = 0|\boldsymbol{v} = \boldsymbol{x}_\alpha) = 1 - \gamma_\alpha$. Furthermore, let β_α be the probability that a background location is detected as a part (or false alarm rate), that is $\beta_\alpha = P(d^{\boldsymbol{v}}_\alpha(I) = 1|\boldsymbol{v} \neq \boldsymbol{x}_\alpha)$. Thus, $P(d^{\boldsymbol{v}}_\alpha(I) = 0|\boldsymbol{v} \neq \boldsymbol{x}_\alpha) = 1 - \beta_\alpha$. In addition, let d_0 and d_1 denote the number of non-responses and responses, respectively. Hence, $d_0 + d_1$ is the size of image I.

We make a further assumption that the responses are independent given the true location of parts. Again, this assumption is not strictly true, because object parts are correlated with neighboring locations. Nevertheless, this assumption reduces the model complexity greatly. The conditional probability distribution $P_\alpha(\cdot\,|\,\cdot)$ in Eq. (3) can be expressed as

$$\begin{aligned} P_\alpha(d_\alpha(I)|\boldsymbol{x}_\alpha) &= P(d_\alpha^{\boldsymbol{v}}(I)|\boldsymbol{v}=\boldsymbol{x}_\alpha)\prod_{\boldsymbol{v}\neq\boldsymbol{x}_\alpha} P(d_\alpha^{\boldsymbol{v}}(I)|\boldsymbol{v}\neq\boldsymbol{x}_\alpha) \\ &= \gamma_\alpha^{d_\alpha^{\boldsymbol{x}_\alpha}(I)}(1-\gamma_\alpha)^{1-d_\alpha^{\boldsymbol{x}_\alpha}(I)}(1-\beta_\alpha)^{d_0-(1-d_\alpha^{\boldsymbol{x}_\alpha}(I))}\beta_\alpha^{d_1-d_\alpha^{\boldsymbol{x}_\alpha}(I)} \\ &= \left(\frac{\gamma_\alpha(1-\beta_\alpha)}{(1-\gamma_\alpha)\beta_\alpha}\right)^{d_\alpha^{\boldsymbol{x}_\alpha}(I)}(1-\gamma_\alpha)(1-\beta_\alpha)^{d_0-1}\beta_\alpha^{d_1} \\ &= \left(\frac{\gamma_\alpha(1-\beta_\alpha)}{(1-\gamma_\alpha)\beta_\alpha}\right)^{d_\alpha^{\boldsymbol{x}_\alpha}(I)} c_\alpha, \end{aligned} \tag{4}$$

where c_α is a value that does not depends on x. Substituting (4) into (3), we have

$$P(\{d_\alpha(I)\}_{1\leqslant\alpha\leqslant p}|x) = \prod_{1\leqslant\alpha\leqslant p}\left(\frac{\gamma_\alpha(1-\beta_\alpha)}{(1-\gamma_\alpha)\beta_\alpha}\right)^{d_\alpha^{\boldsymbol{x}_\alpha}(I)} c_\alpha \tag{5}$$

and in the log form

$$\log(P(\{d_\alpha(I)\}_{1\leqslant\alpha\leqslant p}|x)) = c + \sum_{1\leqslant\alpha\leqslant p} d_\alpha^{\boldsymbol{x}_\alpha}(I)\log\left(\frac{\gamma_\alpha(1-\beta_\alpha)}{(1-\gamma_\alpha)\beta_\alpha}\right), \tag{6}$$

where c is a value that does not depend on x.

An intuitive interpretation of Eq. (6) is that a bonus value of $\log(\frac{\gamma_\alpha(1-\beta_\alpha)}{(1-\gamma_\alpha)\beta_\alpha})$ is awarded if part α is detected. For the method to bias towards detected parts, this value should be greater than zero, which is equivalent to $\gamma_\alpha > \beta_\alpha$. Informally, the detection rate should be greater than the probability that a background location is detected as a part.

In short, Eq. (6) is our approximation of the image likelihood given an object. When used in the objective function to evaluate different values of x, the constant c can be ignored.

3.2. The prior

This section presents a model for the object configuration. First, we present a translation invariant model by moving the object location to a new coordinate centered at an object part. The chosen object part is called the mapping point. We then address the problem when this mapping point is not detected. Finally, we present our solution to the problem of missed detection of other non-mapping points in the configuration.

3.2.1. Translation invariance

Consider p ordered, labeled points $(X_i, Y_i), X_i, Y_i \in \mathcal{R}, i = 1,\ldots,p$. A $2p$-vector $\boldsymbol{Z}$ can be constructed

$$\boldsymbol{Z} = (X_1,\ldots,X_p,Y_1,\ldots,Y_p)'.$$

A new vector $\boldsymbol{Z}^*$ can be obtained from $\boldsymbol{Z}$ by mapping (X_1, Y_1) to the origin

$$\boldsymbol{Z}^* = (X_2 - X_1,\ldots,X_p - X_1, Y_2 - Y_1,\ldots,Y_p - Y_1)'. \tag{7}$$

We assume that the configuration after translation normalization $\boldsymbol{Z}^*$ is jointly Gaussian, denoted by $\boldsymbol{Z}^* \sim \mathcal{N}_{2p-2}(\boldsymbol{\mu}^*, \Sigma)$. We can approximate the density distribution of $\boldsymbol{Z}$ by $\boldsymbol{Z}^*$. Effectively, we have a translation invariant configuration space. The density of $\boldsymbol{Z}^*$ is given by

$$P(\boldsymbol{Z}^*) = (2\pi)^{-\frac{1}{2}(2p-2)}|\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{Z}^* - \boldsymbol{\mu}^*)'\Sigma^{-1}(\boldsymbol{Z}^* - \boldsymbol{\mu}^*)\right] \tag{8}$$

and in the log form

$$\log(P(\boldsymbol{Z}^*)) = k + \left[-\frac{1}{2}(\boldsymbol{Z}^* - \boldsymbol{\mu}^*)'\Sigma^{-1}(\boldsymbol{Z}^* - \boldsymbol{\mu}^*)\right], \tag{9}$$

where k is a value that does not depend on x.

First, we review two properties of a multivariate Gaussian distribution that will be needed.

Lemma 1. *Let $\boldsymbol{t}$ be an ℓ-vector and $\boldsymbol{t} \sim \mathcal{N}_\ell(\boldsymbol{\theta}, \Sigma)$. Then for any $r \times \ell$ matrix C, $C\boldsymbol{t} \sim \mathcal{N}_r(C\boldsymbol{\theta}, C\Sigma C')$.*

Consult [28] for the proof.

Lemma 2. *Let $\boldsymbol{t} = (\boldsymbol{t}_1', \boldsymbol{t}_2')' \sim \mathcal{N}_\ell(\boldsymbol{\theta}, \Sigma)$, that is vector $\boldsymbol{t}$ is decomposed into $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$. Let $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$ be r- and $(\ell - r)$-vectors, respectively. Suppose the corresponding partition of $\boldsymbol{\theta}$ and Σ are, respectively, given by*

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{pmatrix},$$

where $\boldsymbol{\theta}_1$ is an r-vector and Σ_{11} is an $r \times r$ matrix. Then the conditional distribution $P(\boldsymbol{t}_1|\boldsymbol{t}_2)$ is

$$\mathcal{N}_r(\boldsymbol{\theta}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{t}_2 - \boldsymbol{\theta}_2), \Sigma_{1\cdot 2}),$$

where

$$\Sigma_{1\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}'$$

Again, consult [28] for the proof.

So far we have considered the case where all parts are detected. When some parts are missed, we still need to approximate the density. A special case is when the part used as mapping point is missed. First, we show that $P(\boldsymbol{Z}^*)$ can be computed when a different point is used as mapping point. Then we describe a procedure for estimating $P(\boldsymbol{Z}^*)$ when missing values occur.

3.2.2. Mapping point invariance

Consider a situation when point (X_i, Y_i), $i \neq 1$, is mapped to the origin instead of (X_1, Y_1). This is needed in case where the first part is not detected. In that case, a new $(2p-2)$-vector can be constructed

$$\boldsymbol{Z}_i^* = (X_1 - X_i, \ldots, Y_1 - Y_i, \ldots)'. \tag{10}$$

Vector $\boldsymbol{Z}_i^*$ can be obtained from $\boldsymbol{Z}^*$ via a linear transformation, that is $\boldsymbol{Z}_i^* = L_i \boldsymbol{Z}^*$, where the transformation matrix L_i is defined as follows. The entries in the column $(i-1)$ of L_i are (-1)'s. In every other column j, there is only one value of 1 at row $(j+1)$ for $j < (i-1)$, and at row j for $j > (i-1)$. All other entries are 0. An example of a transformation matrix of size 8 is given below

$$L_4^{(8)} = \begin{bmatrix} 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Therefore, by Lemma 1, $\boldsymbol{Z}_i^* \sim \mathcal{N}_{2p-2}(L_i\boldsymbol{\mu}^*, L_i \Sigma L_i')$. Moreover, $P(\boldsymbol{Z}^*)$ is proportional to $P(\boldsymbol{Z}_i^*)$ with a factor equal the absolute value of $\det(L_i)$ which is 1 (the sub-matrix obtained by the elimination of the first row and column $(i-1)$ is an identity matrix). Thus, $P(\boldsymbol{Z}^*) = P(\boldsymbol{Z}_i^*)$.

3.2.3. Approximation under missing parts

When a set of parts is left undetected, we assume that the missing parts are located at their best locations for the sake of the global configuration. Note that given a fixed state of detected and non-detected parts, the likelihood term of Eq. (2) does not change. Therefore, we can maximize $P(\boldsymbol{Z}^*)$.

When the first part is detected, an obvious choice for the mapping point is (X_1, Y_1). Otherwise, because our computation does not depend on the choice of the mapping point, we can map any detected point to the origin. For generality, let (X_o, Y_o) be the mapping point. This results in a vector $\boldsymbol{Z}_o^*$ of $(2p-2)$ dimensions. This is a random vector distributed according to a multivariate Gaussian distribution $\boldsymbol{Z}_o^* \sim \mathcal{N}_{2p-2}(L_o\boldsymbol{\mu}^*, L_o \Sigma L_o')$, where L_o is given in Section 3.2.2.

We can decompose $\boldsymbol{Z}_o^*$ into two parts $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$ as in Lemma 2, where $\boldsymbol{t}_2$ denotes the observed components and $\boldsymbol{t}_1$ denotes the missing parts. Further, we have $P(\boldsymbol{Z}_o^*) = P(\boldsymbol{t}_2)P(\boldsymbol{t}_1|\boldsymbol{t}_2)$. Maximizing $P(\boldsymbol{Z}_o^*)$ is therefore equivalent to maximizing $P(\boldsymbol{t}_1|\boldsymbol{t}_2)$. According to Lemma 2, $P(\boldsymbol{t}_1|\boldsymbol{t}_2)$ is also a multivariate Gaussian distribution; therefore $P(\boldsymbol{t}_1|\boldsymbol{t}_2)$ is maximal when $\boldsymbol{t}_1$ equals the mean of the distribution, which is specified in Lemma 2. Given the value of $\boldsymbol{t}_1$, $P(\boldsymbol{Z}_o^*)$ can be computed easily, which is also the value of $P(\boldsymbol{Z}^*)$.

3.3. Parameter estimation

The final form of the objective function $f(x)$ is obtained by substituting Eqs. (6) and (9) into Eq. (2), ignoring the constant values

$$f(x) = \sum_{1 \leqslant \alpha \leqslant p} d_\alpha^{x_\alpha}(I) \log\left(\frac{\gamma_\alpha(1-\beta_\alpha)}{(1-\gamma_\alpha)\beta_\alpha}\right) + \left[-\frac{1}{2}(\boldsymbol{Z}^* - \boldsymbol{\mu}^*)' \Sigma^{-1} (\boldsymbol{Z}^* - \boldsymbol{\mu}^*)\right], \tag{11}$$

where $\boldsymbol{Z}^*$ is obtained from x as in Eq. (7). Learning the object model involves the estimation of the parameters $\{\gamma_\alpha, \mu_\alpha | 1 \leqslant \alpha \leqslant p\}$, $\boldsymbol{\mu}^*$, and Σ.

When missing part occurs, we optimize $f(x)$ as specified in 3.2.3. Rather than marginalizing over the missing part positions, we consider them being at their optimal locations. This objective function reflects a tradeoff between detection and locations of parts. To illustrate the point, consider a triangular object ABC with two parts A and B detected in Fig. 1. The ideal location of C is indicated in the figure. The objective function favors the detected part C closest to this ideal location. However, if this closest location is still two far (depending on detector performance of this part), it is better to declare a miss because the penalty for geometrical distortion outweighs the bonus for detected part.

One approach to parameter estimation is to label the training data manually for object parts. From the labeled points, we extract image patches to serve as examples for object parts. In addition, we extract randomly patches from the background to serve as non-part examples.

The value of γ_α and β_α can be estimated by applying the detector d_α on the training set, then computing the detection rate and false alarm rate for part α. Alternatively, we can define a certain class of detectors, and subsequently train detector d_α using the training examples for part α. In this case, the values of γ_α and β_α can be estimated in various ways depending on the number of training examples available. One can employs the hold-out method by splitting the dataset further into training set and test set. Another method is n-fold cross validation where the dataset is divided into n approximately equal partitions. The training is carried out n times; each time

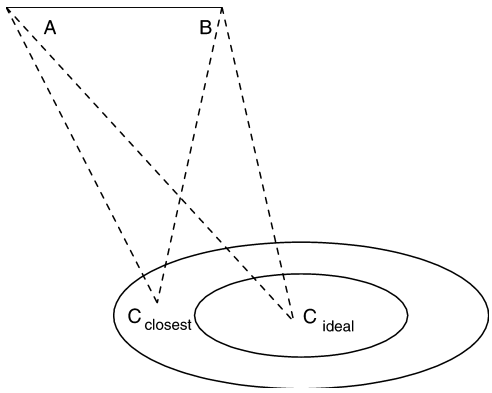

Fig. 1. An example of a triangular object where two parts A and B are detected and part C is uncertain. The objective function favors detected part C closest to the ideal location, but only within certain extent. For a specific covariance, the outside ellipse is the maximum extent for a detector with detection rate (γ) of 0.9 and false alarm (β) of 0.01 (C_{closest} is selected). The inside ellipse denotes the extent of a detector with detection rate (γ) of 0.5 and false alarms (β) of 0.1 (missing part is declared).

$n-1$ different partitions are used for training the one partition is used for testing. The final estimation is the average of the estimation of the n runs.

The parameters for the Gaussian distribution $\boldsymbol{\mu}^*$ and Σ are estimated from the locations of the labeled points. In case there are missing points in the training examples, one can make use of the expectation maximization algorithm [29,30].

4. Searching for the optimal solution

Given Eq. (11) for the objective function, the next topic is how to find the optimal solution efficiently. The geometric part of this objective function has no special structure. Because we impose no constraint on the covariance matrix and its inverse, this term cannot be decomposed into chain-like sum of terms. This makes the optimization of the objective function non-trivial. However, by exploiting two properties of the Gaussian distribution (Lemmas 1 and 2), we propose an efficient optimization method for this objective function with an A^* search. We first review briefly the A^* search algorithm [21]. We then formulate the problem as a graph search problem, and subsequently establish the necessary heuristic to use the A^* search algorithm to find the optimal solution.

The A^* search is an algorithm to find an optimal path from a start vertex A to a destination vertex B in a graph. We consider the maximum score problem. At each vertex C, the algorithm estimates the score $f(C)$ of the path from A to B going through C. The score $f(C)$ is the sum of the score of the real path from A to C and an estimated score going from C to B (see Fig. 2). The algorithm stores all explored sub-paths in a priority queue, and iteratively expands the sub-path with the highest estimated score $f(C)$. The first path reaching B is the solution of the algorithm. It is proved that if the estimated score (heuristic) is always greater than the true score, the algorithm will lead to an optimal solution. The heuristic in that case is called an *admissible* heuristic. The problem is to devise a heuristic for a particular problem.

First we formulate the optimization problem as a graph search problem. For each input image, we construct a directed weighted graph of $(2+p+\sum_{1\leqslant\alpha\leqslant p}M_\alpha)$ nodes as illustrated in Fig. 3, where M_α denotes the number of occurrences of part α. The construction of the graph is as follows. There is one starting vertex and destination vertex denoted by A and B, respectively. For each object part α, there is one vertex $w_{\alpha i}$ for each detected location i, $1\leqslant i\leqslant M_\alpha$. In addition, there is one vertex $w_{\alpha 0}$ denoting the missing value. Every path from A to B specifies an object hypothesis and is

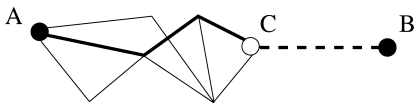

Fig. 2. At each vertex C, an estimate $f(C)$ of the score going from A to B through C guides the exploration of the A^* search algorithm.

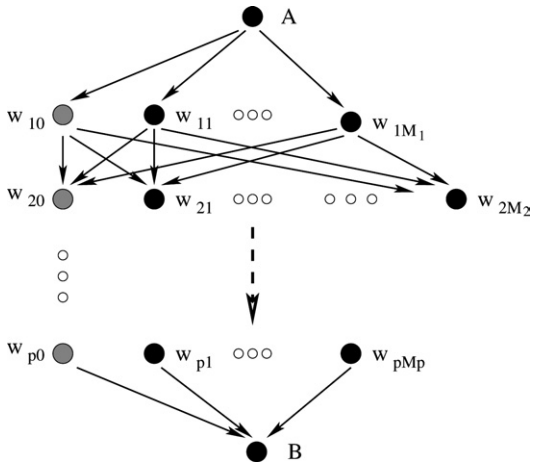

Fig. 3. Search graph. M_α is the number of part α detected, for $1 \leqslant \alpha \leqslant p$.

associated with a score measure. This measure is computed by the objective function (Eq. 11). The problem is to find a path from A to B with the highest score.

There are a total of $\prod_{1\leqslant\alpha\leqslant p}(M_\alpha + 1)$ paths from A to B. This number is typically very large. For example, in a test image in the experiment on real data, the number of nodes in the constructed graph is 106 with 3483648 possible hypotheses. Hence, a brutal force search approach is not feasible for a large number of parts in a cluttered scene.

Fortunately, we can derive an estimate of the total score at each vertex along a path. This estimation can be used as a heuristic in an A^* search for the optimal solution.

The derivation of the estimation is informally as follows. Suppose we are at vertex $w_{\alpha k}$, $0 \leqslant k \leqslant M_\alpha$, after traveling path $q_{A\to w_{\alpha k}}$. The vertices in $q_{A\to w_{\alpha k}}$, except A, determine the "best" possible configuration for the remaining vertices as in 3.2.3. Further, imagine that at every "best" location in the unexplored path, there is a corresponding part. It is clear that the score f^* of this ideal path is an upper bound for any real path from A to B with the sub-path $q_{A\to w_{\alpha k}}$. Thus, this heuristic is admissible.

Definition 1. Consider a path from A to $w_{\alpha k}$. Construct a vector $\boldsymbol{Z}^h$ after translation normalization consisting of two parts $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$, where $\boldsymbol{t}_1$ denotes the missing parts and the unexplored parts, and $\boldsymbol{t}_2$ denotes the detected parts. The value of $\boldsymbol{t}_1$ is given as in Section 3.2.3. The heuristic $f^*(w_{\alpha k})$ to estimate the score to complete the path from A to B is

$$f^*(w_{\alpha k}) = \sum_{1\leqslant i\leqslant\alpha} d_i^{x_i}(I)\log\left(\frac{\gamma_i(1-\beta_i)}{(1-\gamma_i)\beta_i}\right) + \sum_{\alpha<j\leqslant p}\log\left(\frac{\gamma_j(1-\beta_j)}{(1-\gamma_j)\beta_j}\right) + \left[-\frac{1}{2}(\boldsymbol{Z}^h-\boldsymbol{\mu}^*)'\Sigma^{-1}(\boldsymbol{Z}^h-\boldsymbol{\mu}^*)\right]. \quad (12)$$

Lemma 3. *$f^*(w_{\alpha k})$ is admissible.*

Proof. Consider any path completing the travel to B from $w_{\alpha k}$. Let $\boldsymbol{Z}^r$ denotes the vector after translation normalization. The real score is

$$\begin{aligned} f^r(w_{\alpha k}) &= \sum_{1<i\leqslant p} d_i^{\boldsymbol{x}_i}(I)\log\left(\frac{\gamma_i(1-\beta_i)}{(1-\gamma_i)\beta_i}\right) + \left[-\frac{1}{2}(\boldsymbol{Z}^r-\boldsymbol{\mu}^*)'\Sigma^{-1}(\boldsymbol{Z}^r-\boldsymbol{\mu}^*)\right] \\ &= \sum_{1\leqslant i\leqslant \alpha} d_i^{\boldsymbol{x}_i}(I)\log\left(\frac{\gamma_i(1-\beta_i)}{(1-\gamma_i)\beta_i}\right) + \sum_{\alpha<j\leqslant p} d_j^{\boldsymbol{x}_j}(I)\log\left(\frac{\gamma_j(1-\beta_j)}{(1-\gamma_j)\beta_j}\right) \\ &\quad + \left[-\frac{1}{2}(\boldsymbol{Z}^r-\boldsymbol{\mu}^*)'\Sigma^{-1}(\boldsymbol{Z}^r-\boldsymbol{\mu}^*)\right]. \end{aligned} \tag{13}$$

We have that the first sum of the true score is identical to that of the heuristic. As for the second sum, recall that $d_i^{\boldsymbol{x}_i}(I)$ is a binary value, hence

$$\sum_{\alpha<i\leqslant p}\log\left(\frac{\gamma_i(1-\beta_i)}{(1-\gamma_i)\beta_i}\right) \geqslant \sum_{\alpha<i\leqslant p} d_i^{\boldsymbol{x}_i}(I)\log\left(\frac{\gamma_i(1-\beta_i)}{(1-\gamma_i)\beta_i}\right)$$

As for the final term, $\boldsymbol{Z}^r$ can be decomposed into two component $\boldsymbol{t}_1^r$ and $\boldsymbol{t}_2$ where $\boldsymbol{t}_2$ is the detected components as in Definition 1. This is possible because the path from A to $w_{\alpha k}$ is fixed. As specified in Section 3.2.3, the value of $\boldsymbol{t}_1$ in Definition 1 is optimal for the configuration given the observed value $\boldsymbol{t}_2$. Hence, the geometrical part of the heuristic is not less than that of the true score.

Therefore, the heuristic $f^*(w_{\alpha k})$ always overestimates the true score, and hence it is admissible. □

The A^* search will explore the vertex with highest f^* value. This strategy guarantees that the first path exploring B is an optimal path [21].

5. Experiments

5.1. On synthetic data

The purpose of this experiment is to show the working of the objective function as well as the efficiency of the search procedure.

The performance of the A^* search is often judged by the number of sub-paths that are in the search queue when the solution is reached. A small value for this number indicates that the solution is reached quickly [21]. This value and the total number of possible paths (of a brutal force search) will be used to evaluate the algorithm.

First, we create 10,000 samples of human figures. Each object consists of 11 parts. The creation of objects is as follows. Basically, each object forms a tree, as shown in Fig. 4. We fix the root of the tree, and the subsequent nodes are generated according to pre-defined Gaussian distributions conditional on the previous node. Fig. 4 shows fifteen examples of these figures.

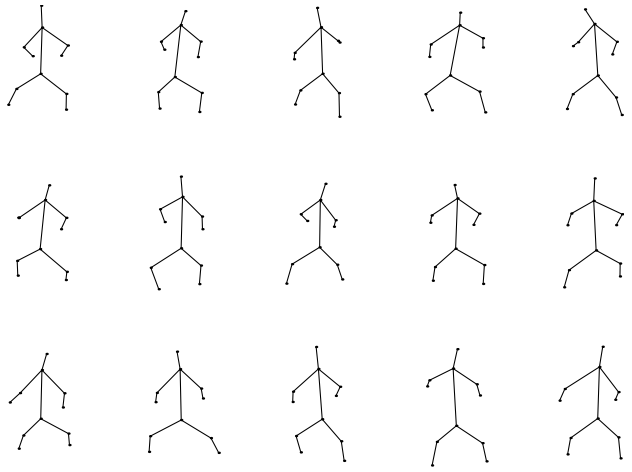

Fig. 4. Fifteen examples of a human figure; each is an object consisting of 11 parts.

Each object is then placed on an image of size 300×300 at a random position. For each pair of image and object, we generate two types of noise as discussed in Section 3.1. For each part, the number of false alarms is drawn from a Poisson distribution with a pre-defined mean.

The search method in Section 4 is applied to find an object in each image.

An example of the part detected image is given in Fig. 5A. The corresponding search result is shown in Fig. 5B. The method is able to locate objects in case of

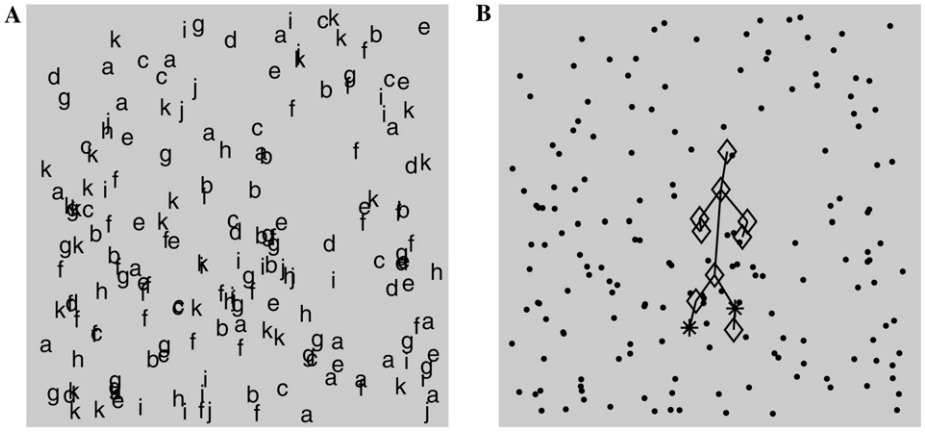

Fig. 5. An example where the object is correctly located in case of two missing parts denoted by asterisks (*). (A) Part detection results. Parts are denoted by letters from a to k. The total number of possible paths is close to 10^{14}. (B) Search result. The total number of sub-paths in the queue when the optimal solution is reached is 2437 only.

missing parts and with the presence of clutter. When missing parts occur, the method also approximates their locations.

The number of sub-paths that are in the search queue when the solution is reached is very small in comparison with the total number of paths. This shows an excellent performance of the search algorithm.

5.2. On real images

We use the Caltech face dataset [13]. The dataset consists of 450 images, each containing a human face in a clutter background with various lighting conditions. Upon examination of the dataset, we remove nine images (from image 328 up to 336) due to the small scale and three hand-draw images (image 400, 402, and 403). These images are significantly different from the rest of the dataset. In total, there are 438 images of 28 people. The faces appear at different locations in each image with a variety of facial expressions. The original images are converted into grayscale images and downsampled by half to a resolution of 448×296 pixels.

We split the dataset into two partitions. The training set contains 216 images of 14 people. The test set contains 222 images of the remaining 14 people. For each training image, we manually labeled five object parts: the two eyes, the nose, and the two corners of the mouth, each by one mouse click. The size of each part is 32×32 pixels. There is no evidence why this decomposition is optimal, but for the purpose of this paper, the choice is sufficient.

We use a simple form of object part detectors, namely template matching with normalized cross correlation [31]. Given an input template u, the normalized cross correlation ρ between object template t_α and u is defined as

$$\rho(t_\alpha, u) = \frac{\sum_{\boldsymbol{i}} (t_\alpha(\boldsymbol{i}) - \bar{t}_\alpha)(u(\boldsymbol{i}) - \bar{u})}{\left(\sum_{\boldsymbol{i}} (t_\alpha(\boldsymbol{i}) - \bar{t}_\alpha)^2 \sum_{\boldsymbol{i}} (u(\boldsymbol{i}) - \bar{u})^2 \right)^{1/2}}, \tag{14}$$

where $\boldsymbol{i}$ denotes the two-dimensional index vector, $t_\alpha(\boldsymbol{i})$ and $u(\boldsymbol{i})$ denote the pixel values of the templates at image location $\boldsymbol{i}$, and $\bar{t}_\alpha$ and $\bar{u}$ denote the average pixel values of the two templates. The template t_α for each part α is the average of the training examples aligned by the manually selected centers. Fig. 6 shows the five templates obtained, as arranged according to the mean of the configuration model. An object part is detected by applying its detector at all image locations, followed by a non-max suppression operation on the responses.

A threshold value is needed to determine whether an image patch u is classified as object part α. In this experiment, we set the false alarm rate for each detector at approximately 10^{-2}. To estimate the false alarm rate, we sample uniformly 20,000 patches of size 32×32 from the training images. The false alarm rate is rather unimportant for the purpose of this paper since we want to show that the method works for various detectors and various uncertainties of the part detection. (We note that we expect a correlation in the performance of the two eye detectors, distorting the independence assumption.)

Fig. 6. Templates of the five selected parts of a human face. The templates are arranged according to the average configuration.

Fig. 7 shows the detection result of four images in the test set. The faces are correctly localized in a cluttered scene and there is some missed detection. Overall, the method obtains a correct localization rate of 92% on the training set and 91% on the test set. Fig. 8 shows two examples where the faces are not localized correctly. In general, mistake happens in cases where the detection of many parts fails due to

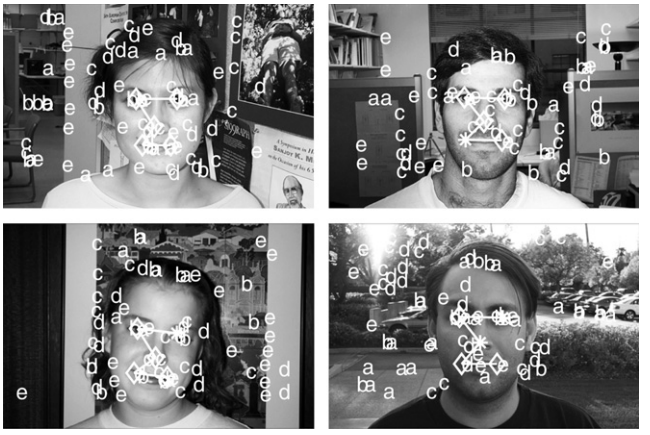

Fig. 7. Example of the detection results. The search algorithm is able to locate the object in a cluttered scene. It also works in case missed detection occurs. The asterisks (*) indicate the likely location of the missed parts. The letters from a to e denote object parts.

Fig. 8. Example of incorrect object localization. In both cases the part detectors fail to find a reasonable number of correct parts. In the image on the left only the left eye (denoted by letter b) is located correctly. In the image on the right, only the nose (denoted by letter c) is located correctly.

non-standard lighting conditions. Again, this is not the purpose of this paper. This paper aims to lay down a method for spatially combining classifiers.

The recognition result depends on the composition of object parts. In our experiment (data not shown), we are able to raise the correct localization rate from 91 to 96.5% on the test set using a more advanced part-labeling scheme. Thus, we consider automatic part learning is an important track of research.

It is difficult to compare this result with the result of Burl and Perona [2] because they use a different dataset. They obtain 94% correct localization on a dataset of 130 images of one person. However, on the training set with a variety of individuals, the result is only 63%. Note that our test set consists of 14 individuals and is different from the training set.

On average, the search space consists of 556848 hypotheses, but the search reaches the optimal solution when there are about 360 sub-paths in the queue only. This translates into approximately half of a second on a 1 GHz CPU with a simple Matlab implementation. The main computational cost is the detection of the five object parts, which take a few seconds.

6. Discussion and conclusion

In this paper we consider the recognition of an object once its parts have been detected. The problem is how to encode the global configuration of an object and the presence of its parts in a cluttered scene into an objective function in order to evaluate object hypotheses. In addition, an efficient search procedure is required for finding the best hypothesis according to the objective function. This is because a potentially large search space is anticipated.

We propose an objective function in the Bayesian framework. This criterion is capable of encoding the global configuration, as well as information of the cluttered scene using the performance statistics of the part detectors. An important characteristic of the proposed objective function is that it allows a fast search algorithm, namely the A^* search. By deriving an upper bound on the final value of the objective function, we effectively achieve an *admissible* heuristic for the A^* search, which guarantees an optimal solution.

We performed experiments on both synthetic and real data. The experimental results show an excellent performance of the proposed heuristic for the search problem. The high recognition rate achieved also indicates that the objective function is suitable model for the problem of recognition by parts.

Our recognition system is placed in a unified probabilistic framework. The probabilistic model shows an interesting link between the recognition of objects and performance of part detectors, detection algorithms and the cluttered scene. The future research will focus on the recognition performance with respect to the performance of the part detectors. This includes investigation into the independence assumption among part detectors, the use of likelihood values returned by detectors when available, and the combination of part detection and the statistics of natural images.

For a robust recognition system, the problems of self-occlusion and natural occlusion need to be addressed, together with the study of various scene accidental conditions. We also leave that for future investigation.

In conclusion, we propose a probabilistic framework for the object recognition problem. The two important components of this framework are the objective function and the search method. The objective function is derived in a Bayesian approach, handling of the uncertainty of the part detection. It is able to encode the object global configuration as well as the presence of clutter. In addition, a heuristic is devised for the A^* search, which achieves an optimal solution for recognition.

Acknowledgments

We thank the reviewers for their helpful comments, which have strengthened the paper. This research is sponsored by the ICES/KIS Multimedia Information Analysis project, TNO Institute of Applied Physics, and MultimediaN.

References

[1] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cognit. Neuro Sci. 3 (1) (1991) 71–86.
[2] M. Burl, P. Perona, Recognition of planar object classes, in: Proc. of CVPR'96, 1996, pp. 223–230.
[3] K.K. Sung, T. Poggio, Example-based learning for view-based human face detection, IEEE PAMI 20 (1) (1998) 39–51.
[4] H.A. Rowley, S. Baluja, T. Kanade, Neural network-based face detection, IEEE PAMI 20 (1) (1998) 23–38.
[5] S. Agarwal, D. Roth, Learning a sparse representation for object detection, in: Proc. of ECCV, 2002, pp. 113–130.
[6] W.E.L. Grimson, Object Recognition by Computer: The Role of Geometric Constraints, MIT Press, Cambridge, MA, 1990.
[7] F.L. Bookstein, Size and shape spaces for landmark data in two dimensions (with discussion), Stat. Sci. 1 (2) (1986) 181–242.
[8] D.G. Kendall, A survey of the statistical theory of shape, Stat. Sci. 4 (2) (1989) 87–120.
[9] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active shape models-their training and application, Comput. Vis. Image Understand. 61 (1995) 38–59.
[10] S. Ghebreab, A. Smeulders., Strings: variational deformable models of multivariate ordered features, IEEE PAMI 25 (11) (2003) 1–14.
[11] E. Klassen, A. Srivastava, W. Mio, S.H. Joshi, Analysis of planar shapes using geodesic paths on shape spaces, IEEE PAMI 26 (3) (2004) 372–383.
[12] I.L. Dryden, K.V. Mardia, General shape distributions in a plane, Adv. Appl. Probab. 23 (1991) 259–276.
[13] M. Weber, M. Welling, P. Perona, Unsupervised learning of models for recognition, in: Proc. of ECCV, 2000, pp. 18–32.
[14] L. Staib, J. Duncan, Boundary finding with parametrically deformable models, IEEE PAMI 14 (11) (1992) 1061–1075.
[15] J. MacCormick, A. Blake, A probabilistic contour discriminant for object localisation, in: Proc. 6th ICCV, 1998, pp. 390–395.
[16] J. Sullivan, A. Blake, M. Isard, J. MacCormick, Bayesian object localisation in images, Int. J. Comput. Vis. 44 (2) (2001) 111–135.

[17] A.J. Baddeley, M.N.M. van Lieshout, Recognition of overlapping objects using Markov spatial processes, Tech. rep., CWI, bS-R9109, ISSN 0924-0659, 1991.
[18] W.M. Wells, Statistical approaches to feature-based object recognition, Inter. J. Comput. Vis. 21 (1/2) (1997) 63–98.
[19] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient matching of pictorial structures, in: Proc. of CVPR'2000, vol. 2, 2000, pp. 66–73.
[20] J.M. Coughlan, D. Snow, C. English, A.L. Yuille, Efficient deformable template detection and localization without user initialization, Comput. Vis. Image Understand. 78 (2000) 303–319.
[21] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, Prentice Hall, Englewood Cliffs, NJ, 1995.
[22] S. Geman, D. Geman, Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images, IEEE PAMI 6 (6) (1984) 721–741.
[23] M.I. Miller, U. Grenander, J.A. O'Sullivan, D.L. Snyder, Automatic target recognition organized via jump-diffusion algorithms, IEEE Trans. Image Process. 6 (1) (1997) 157–174.
[24] A.L. Yuille, J. Coughlan, An A^* perspective on deterministic optimization for deformable templates, Pattern Recognit. 33 (4) (2000) 603–616.
[25] D. Geman, B. Jedynak, An active testing model for tracking roads in satellite images, IEEE PAMI 18 (1) (1996) 1–14.
[26] S. Zhu, D. Mumford, Prior learning and Gibbs reaction-diffusion, IEEE PAMI 19 (11) (1997) 1236–1250.
[27] U. Grenander, A. Srivastava, Probability models for clutter in natural images, IEEE PAMI 23 (4) (2001) 424–429.
[28] M.S. Srivastava, C.G. Khatri, An introduction to multivariate statistics, North Holland, 1979.
[29] A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Series B 39 (1) (1977) 1–38.
[30] J.L. Schafer, Analysis of Incomplete Multivariate Data, Chapman & Hall, London, 1997.
[31] W.K. Pratt, Digital Image Processing, second ed., Wiley, New York, 1991.