WIKIPEDIA

Constellation model

The **constellation model** is a probabilistic, generative model for category-level object recognition in computer vision. Like other part-based models, the constellation model attempts to represent an object class by a set of *N* parts under mutual geometric constraints. Because it considers the geometric relationship between different parts, the constellation model differs significantly from appearance-only, or "bag-of-words" representation models, which explicitly disregard the location of image features.

The problem of defining a generative model for object recognition is difficult. The task becomes significantly complicated by factors such as background clutter, occlusion, and variations in viewpoint, illumination, and scale. Ideally, we would like the particular representation we choose to be robust to as many of these factors as possible.

In category-level recognition, the problem is even more challenging because of the fundamental problem of intra-class variation. Even if two objects belong to the same visual category, their appearances may be significantly different. However, for structured objects such as cars, bicycles, and people, separate instances of objects from the same category are subject to similar geometric constraints. For this reason, particular parts of an object such as the headlights or tires of a car still have consistent appearances and relative positions. The Constellation Model takes advantage of this fact by explicitly modeling the relative location, relative scale, and appearance of these parts for a particular object category. Model parameters are estimated using an unsupervised learning algorithm, meaning that the visual concept of an object class can be extracted from an unlabeled set of training images, even if that set contains "junk" images or instances of objects from multiple categories. It can also account for the absence of model parts due to appearance variability, occlusion, clutter, or detector error.

Contents

History

The idea for a "parts and structure" model was originally introduced by Fischler and Elschlager in 1973.[1] This model has since been built upon and extended in many directions. The Constellation Model, as introduced by Dr. Perona and his colleagues, was a probabilistic adaptation of this approach.

In the late '90s, Burl et al.[2][3][4][5] revisited the Fischler and Elschlager model for the purpose of face recognition. In their work, Burl et al. used manual selection of constellation parts in training images to construct a statistical model for a set of detectors and the relative locations at which they should be applied. In 2000, Weber et al. [6][7][8][9] made the significant step of training the model using a more unsupervised learning process, which precluded the necessity for tedious hand-labeling of parts. Their algorithm was particularly remarkable because it performed well even on cluttered and occluded image data. Fergus et al.[10][11] then improved upon this model by making the learning step fully unsupervised, having both shape and appearance learned simultaneously, and accounting explicitly for the relative scale of parts.

The method of Weber and Welling et al.[9]

In the first step, a standard interest point detection method, such as Harris corner detection, is used to generate interest points. Image features generated from the vicinity of these points are then clustered using k-means or another appropriate algorithm. In this process of vector quantization, one can think of the centroids of these clusters as being representative of the appearance of distinctive object parts. Appropriate feature detectors are then trained using these clusters, which can be used to obtain a set of candidate parts from images.

As a result of this process, each image can now be represented as a set of parts. Each part has a type, corresponding to one of the aforementioned appearance clusters, as well as a location in the image space.

Basic generative model

Weber & Welling here introduce the concept of *foreground* and *background*. *Foreground* parts correspond to an instance of a target object class, whereas *background* parts correspond to background clutter or false detections.

Let T be the number of different types of parts. The positions of all parts extracted from an image can then be represented in the following "matrix,"

$$X^o = \begin{pmatrix} x_{11}, x_{12}, \cdots, x_{1N_1} \\ x_{21}, x_{22}, \cdots, x_{2N_2} \\ \vdots \\ x_{T1}, x_{T2}, \cdots, x_{TN_T} \end{pmatrix}$$

where N_i represents the number of parts of type $i \in \{1, \dots, T\}$ observed in the image. The superscript o indicates that these positions are *observable*, as opposed to *missing*. The positions of unobserved object parts can be represented by the vector x^m . Suppose that the object will be composed of F distinct foreground parts. For notational simplicity, we assume here that $F = T$, though the model can be generalized to $F > T$. A *hypothesis* h is then defined as a set of indices, with $h_i = j$, indicating that point x_{ij} is a foreground point in X^o . The generative probabilistic model is defined through the joint probability density $p(X^o, x^m, h)$.

Model details

The rest of this section summarizes the details of Weber & Welling's model for a single component model. The formulas for multiple component models[8] are extensions of those described here.

To parametrize the joint probability density, Weber & Welling introduce the auxiliary variables b and n , where b is a binary vector encoding the presence/absence of parts in detection ($b_i = 1$ if $h_i > 0$, otherwise $b_i = 0$), and n is a vector where n_i denotes the number of *background* candidates included in the i^{th} row of X^o . Since b and n are completely determined by h and the size of X^o , we have $p(X^o, x^m, h) = p(X^o, x^m, h, n, b)$. By decomposition,

$$p(X^o, x^m, h, n, b) = p(X^o, x^m | h, n, b) p(h | n, b) p(n) p(b)$$

The probability density over the number of background detections can be modeled by a Poisson distribution,

$$p(n) = \prod_{i=1}^{T} \frac{1}{n_i!} (M_i)^{n_i} e^{-M_i}$$

where M_i is the average number of background detections of type i per image.

Depending on the number of parts F , the probability $p(b)$ can be modeled either as an explicit table of length

2^F , or, if F is large, as F independent probabilities, each governing the presence of an individual part.

The density $p(h|n,b)$ is modeled by

$$p(h|n,b) = \begin{cases} \frac{1}{\prod_{f=1}^{F} N_f^{b_f}}, & \text{if } h \in H(b,n) \\ 0, & \text{for other } h \end{cases}$$

where $H(b,n)$ denotes the set of all hypotheses consistent with b and n, and N_f denotes the total number of detections of parts of type f. This expresses the fact that all consistent hypotheses, of which there are $\prod_{f=1}^{F} N_f^{b_f}$, are equally likely in the absence of information on part locations.

And finally,

$$p(X^o, x^m|h,n) = p_{fg}(z)p_{bg}(x_{bg})$$

where $z = (x^o x^m)$ are the coordinates of all foreground detections, observed and missing, and x_{bg} represents the coordinates of the background detections. Note that foreground detections are assumed to be independent of the background. $p_{fg}(z)$ is modeled as a joint Gaussian with mean μ and covariance Σ.

Classification

The ultimate objective of this model is to classify images into classes "object present" (class C_1) and "object absent" (class C_0) given the observation X^o . To accomplish this, Weber & Welling run part detectors from the learning step exhaustively over the image, examining different combinations of detections. If occlusion is considered, then combinations with missing detections are also permitted. The goal is then to select the class with maximum a posteriori probability, by considering the ratio

$$\frac{p(C_1|X^o)}{p(C_0|X^o)} \propto \frac{\sum_h p(X^o, h|C_1)}{p(X^o, h_0|C_0)}$$

where h_0 denotes the null hypothesis, which explains all parts as background noise. In the numerator, the sum includes all hypotheses, including the null hypothesis, whereas in the denominator, the only hypothesis consistent with the absence of an object is the null hypothesis. In practice, some threshold can be defined such that, if the ratio exceeds that threshold, we then consider an instance of an object to be detected.

Model learning

After the preliminary step of interest point detection, feature generation and clustering, we have a large set of candidate parts over the training images. To learn the model, Weber & Welling first perform a greedy search over possible model configurations, or equivalently, over potential subsets of the candidate parts. This is done in an iterative fashion, starting with random selection. At subsequent iterations, parts in the model are randomly substituted, the model parameters are estimated, and the performance is assessed. The process is complete when further model performance improvements are no longer possible.

At each iteration, the model parameters

$$\Theta = \{\mu, \Sigma, p(b), M\}$$

are estimated using expectation maximization. μ and Σ, we recall, are the mean and covariance of the joint Gaussian $p_{fg}(z)$, $p(b)$ is the probability distribution governing the binary presence/absence of parts, and M is the mean number of background detections over part types.

M-step

EM proceeds by maximizing the likelihood of the observed data,

$$L(X^o|\Theta) = \sum_{i=1}^{I} \log \sum_{h_i} \int p(X_i^o, x_i^m, h_i|\Theta) dx_i^m$$

with respect to the model parameters Θ. Since this is difficult to achieve analytically, EM iteratively maximizes a sequence of cost functions,

$$Q(\tilde{\Theta}|\Theta) = \sum_{i=1}^{I} E[\log p(X_i^o, x_i^m, h_i|\tilde{\Theta})]$$

Taking the derivative of this with respect to the parameters and equating to zero produces the update rules:

$$\tilde{\mu} = \frac{1}{I} \sum_{i=1}^{I} E[z_i]$$

$$\tilde{\Sigma} = \frac{1}{I} \sum_{i=1}^{I} E[z_i z_i^T] - \tilde{\mu}\tilde{\mu}^T$$

$$\tilde{p}(\bar{b}) = \frac{1}{I} \sum_{i=1}^{I} E[\delta_{b,\bar{b}}]$$

$$\tilde{M} = \frac{1}{I} \sum_{i=1}^{I} E[n_i]$$

E-step

The update rules in the M-step are expressed in terms of sufficient statistics, $E[z]$, $E[zz^T]$, $E[\delta_{b,\bar{b}}]$ and $E[n]$, which are calculated in the E-step by considering the posterior density:

$$p(h_i, x_i^m | X_i^o, \Theta) = \frac{p(h_i, x_i^m, X_i^o | \Theta)}{\sum_{h_i \in H_b} \int p(h_i, x_i^m, X_i^o | \Theta) dx_i^m}$$

The method of Fergus et al.[10]

In Weber et al., shape and appearance models are constructed separately. Once the set of candidate parts had been selected, shape is learned independently of appearance. The innovation of Fergus et al. is to learn not only two, but three model parameters simultaneously: shape, appearance, and relative scale. Each of these parameters are represented by Gaussian densities.

Feature representation

Whereas the preliminary step in the Weber et al. method is to search for the locations of interest points, Fergus et al. use the detector of Kadir and Brady[12] to find salient regions in the image over both location (center) and scale (radius). Thus, in addition to location information X this method also extracts associated scale information S. Fergus et al. then normalize the squares bounding these circular regions to 11 x 11 pixel patches, or equivalently, 121-dimensional vectors in the appearance space. These are then reduced to 10-15 dimensions by principal component analysis, giving the appearance information A.

Model structure

Given a particular object class model with parameters Θ, we must decide whether or not a new image contains an instance of that class. This is accomplished by making a Bayesian decision,

$$\begin{aligned} R &= \frac{p(\text{Object}|X,S,A)}{p(\text{No object}|X,S,A)} \\ &= \frac{p(X,S,A|\text{Object})p(\text{Object})}{p(X,S,A|\text{No object})p(\text{No object})} \\ &\approx \frac{p(X,S,A|\Theta)p(\text{Object})}{p(X,S,A|\Theta_{bg})p(\text{No object})} \end{aligned}$$

where Θ_{bg} is the background model. This ratio is compared to a threshold T to determine object presence/absence.

The likelihoods are factored as follows:

$$p(X,S,A|\Theta) = \sum_{h\in H} p(X,S,A,h|\Theta) =$$

$$\sum_{h\in H} \underbrace{p(A|X,S,h,\Theta)}_{\text{Appearance}} \underbrace{p(X|S,h,\Theta)}_{\text{Shape}} \underbrace{p(S|h,\Theta)}_{\text{Rel. Scale}} \underbrace{p(h|\Theta)}_{\text{Other}}$$

Appearance

Each part p has an appearance modeled by a Gaussian density in the appearance space, with mean and covariance parameters $\Theta_p^{app} = \{c_p, V_p\}$, independent of other parts' densities. The background model has parameters $\Theta_{bg}^{app} = \{c_{bg}, V_{bg}\}$. Fergus et al. assume that, given detected features, the position and appearance of those features are independent. Thus, $p(A|X,S,h,\Theta) = p(A|h,\Theta)$. The ratio of the appearance terms reduces to

$$\frac{p(A|X,S,h,\Theta)}{p(A|X,S,h,\Theta_{bg})} = \frac{p(A|h,\Theta)}{p(A|h,\Theta_{bg})}$$

$$= \prod_{p=1}^{P} \left(\frac{G(A(h_p)|c_p, V_p)}{G(A(h_p)|c_{bg}, V_{bg})} \right)^{b_p}$$

Recall from Weber et al. that h is the hypothesis for the indices of foreground parts, and b is the binary vector giving the occlusion state of each part in the hypothesis.

Shape

Shape is represented by a joint Gaussian density of part locations within a particular hypothesis, after those parts have been transformed into a scale-invariant space. This transformation precludes the need to perform an exhaustive search over scale. The Gaussian density has parameters $\Theta^{\text{shape}} = \{\mu, \Sigma\}$. The background model Θ_{bg} is assumed to be a uniform distribution over the image, which has area α. Letting f be the number of foreground parts,

$$\frac{p(X|S,h,\Theta)}{p(X|S,h,\Theta_{bg})} = G(X(h)|\mu,\Sigma)\alpha^f$$

Relative scale

The scale of each part p relative to a reference frame is modeled by a Gaussian density with parameters $\Theta^{\text{scale}} = \{t_p, U_p\}$. Each part is assumed to be independent of other parts. The background model Θ_{bg} assumes a uniform distribution over scale, within a range r.

$$\frac{p(S|h,\Theta)}{p(S|h,\Theta_{bg})} = \prod_{p=1}^{P} G(S(h_p)|t_p, U_p)^{d_p} r^f$$

Occlusion and statistics of feature detection

$$\frac{p(h|\Theta)}{p(h|\Theta_{bg})} = \frac{p_{\text{Poiss}}(n|M)}{p_{\text{Poiss}}(N|M)} \frac{1}{{}^nC_r(N,f)} p(b|\Theta)$$

The first factor models the number of features detected using a Poisson distribution, which has mean M. The second factor serves as a "book-keeping" factor for the hypothesis variable. The last factor is a probability table for all possible occlusion patterns.

Learning

The task of learning the model parameters $\Theta = \{\mu, \Sigma, c, V, M, p(b|\Theta), t, U\}$ is accomplished by expectation maximization. This is carried out in a spirit similar to that of Weber et al. Details and formulas for the E-step and M-step can be seen in the literature.[11]

Performance

The Constellation Model as conceived by Fergus et al. achieves successful categorization rates consistently above 90% on large datasets of motorbikes, faces, airplanes, and spotted cats.[13] For each of these datasets, the Constellation Model is able to capture the "essence" of the object class in terms of appearance and/or shape. For example, face and motorbike datasets generate very tight shape models because objects in those categories have very well-defined structure, whereas spotted cats vary significantly in pose, but have a very distinctive spotted appearance. Thus, the model succeeds in both cases. It is important to note that the Constellation Model does not generally account for significant changes in orientation. Thus, if the model is trained on images of horizontal airplanes, it will not perform well on, for instance, images of vertically oriented planes unless the model is extended to account for this sort of rotation explicitly.

In terms of computational complexity, the Constellation Model is very expensive. If N is the number of feature detections in the image, and P the number of parts in the object model, then the hypothesis space H is $O(N^P)$. Because the computation of sufficient statistics in the E-step of expectation maximization necessitates evaluating the likelihood for every hypothesis, learning becomes a major bottleneck operation. For this reason, only values of $P \leq 6$ have been used in practical applications, and the number of feature detections N is usually kept within the range of about 20-30 per image.

Variations

One variation that attempts to reduce complexity is the star model proposed by Fergus et al.[14] The reduced dependencies of this model allows for learning in $O(N^2P)$ time instead of $O(N^P)$. This allows for a greater number of model parts and image features to be used in training. Because the star model has fewer parameters, it is also better at avoiding the problem of over-fitting when trained on fewer images.

References

1. M. Fischler and R. Elschlager. *The Representation and Matching of Pictoral Structures.* (1973) (http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1672195)
2. M. Burl, T. Leung, and P. Perona. *Face Localization via Shape Statistics.* (1995) (ftp://vision.caltech.edu/pub/tech-reports-vision/IWAFGR95.ps.Z)
3. T. Leung, M. Burl, and P. Perona. *Finding Faces in Cluttered Scenes Using Random Labeled Graph Matching.* (1995) (ftp://vision.caltech.edu/pub/tech-reports-vision/ICCV95-faces.ps.Z)
4. M. Burl and P. Perona. *Recognition of Planar Object Classes* (1996) (ftp://vision.caltech.edu/pub/tech-reports-vision/CVPR96-recog.ps.gz)
5. M. Burl, M. Weber, and P. Perona. *A Probabilistic Approach to Object Recognition Using Local Photometry and Global Geometry* (1998) (http://www.vision.caltech.edu/publications/ECCV98-recog.pdf)
6. M. Weber. *Unsupervised Learning of Models for Object Recognition.* PhD Thesis. (2000) (http://www.vision.caltech.edu/publications/MarkusWeber-thesis.pdf)
7. M. Weber, W. Einhaeuser, M. Welling and P. Perona. *Viewpoint-Invariant Learning and Detection of Human Heads.* (2000) (ftp://vision.caltech.edu/pub/tech-reports/FG00-recog.pdf)
8. M. Weber, M. Welling, and P. Perona. *Towards Automatic Discovery of Object Categories.* (2000) (ftp://vision.caltech.edu/pub/tech-reports/CVPR00-recog.pdf)
9. M. Weber, M. Welling and P. Perona. *Unsupervised Learning of Models for Recognition.* (2000) (ftp://vision.caltech.edu/pub/tech-reports/ECCV00-recog.pdf)
10. R. Fergus, P. Perona, and A. Zisserman. *Object Class Recognition by Unsupervised Scale-Invariant Learning.* (2003) (ftp://vision.caltech.edu/pub/tech-reports/Fergus_CVPR03.pdf)
11. R. Fergus. *Visual Object Category Recognition.* PhD Thesis. (2005) (http://cs.nyu.edu/~fergus/papers/fergus_thesis.pdf)
12. T. Kadir and M. Brady. *Saliency, scale and image description.* (2001) (http://www.springerlink.com/content/t45n2g8543574026/fulltext.pdf)
13. R. Fergus and P. Perona. Caltech Object Category datasets. http://www.vision.caltech.edu/html-files/archive.html (2003) (http://www.vision.caltech.edu/html-files/archive.html)
14. R. Fergus, P. Perona, and A. Zisserman. *A Sparse Object Category Model for Efficient Learning and Exhaustive Recognition.* (2005) (http://www.robots.ox.ac.uk/%7Efergus/papers/fergus_cvpr05.pdf)

External links

- L. Fei-fei. *Object categorization: the constellation models*. Lecture Slides. (2005) (https://web.archive.org/web/20060912043325/http://courses.ece.uiuc.edu/ece598/ffl/lecture7_ConstellationModel_shortversion.pdf) (link not working)

See also

- Part-based models
- One-shot learning

Retrieved from "https://en.wikipedia.org/w/index.php?title=Constellation_model&oldid=866931027"

This page was last edited on 2 November 2018, at 14:12 (UTC).