Probabilistic Combination of Visual Cues for Object Classification

Roman Filipovych and Eraldo Ribeiro

Computer Vision and Bio-Inspired Computing Laboratory
Department of Computer Sciences
Florida Institute of Technology
Melbourne, FL 32901, USA
rfilipov,eribeiro@fit.edu
http://www.cs.fit.edu/~eribeiro

Abstract. Recent solutions to object classification have focused on the decomposition of objects into representative parts. However, the vast majority of these methods are based on single visual cue measurements. Psychophysical evidence suggests that humans use multiple visual cues to accomplish recognition. In this paper, we address the problem of integrating multiple visual information for object recognition. Our contribution in this paper is twofold. First, we describe a new probabilistic integration model of multiple visual cues at different spatial locations across the image. Secondly, we use the cue integration framework to classify images of objects by combining two-dimensional and three-dimensional visual cues. Classification results obtained using the method are promising.

1 Introduction

The classification (and categorization) of objects and scenes from visual information is one of the most challenging problems in computer vision. Recent advances in image-based object classification have focused on statistical approaches modeling both the appearance of discriminative object parts [3,8] and the spatial relationship among these parts [6,8]. Such methods represent the state-of-the-art in both general object classification and object categorization. However, most object classification methods rely on measurements obtained from a single visual cue. While these methods work remarkably well for specific classes of images, they assume that information about a specific cue is always available. On the other hand, psychophysical evidence suggests that natural vision-based classification tasks are performed better when multiple visual cues can be combined to help reduce ambiguity [14].

In this paper, we address the problem of integrating multiple visual cues using a probabilistic framework. Our contribution is twofold. First, we describe a new model for the integration of a set of distinct visual cues using a Bayesian framework. We model both the cues' appearance information and their spatial structure. As a result, our model allows us to determine the different contributions of each cue in the classification process at different spatial locations in the

G. Bebis et al. (Eds.): ISVC 2007, Part I, LNCS 4841, pp. 662–671, 2007.

object. Secondly, we use our integration framework to classify images of objects by combining two-dimensional and three-dimensional visual cues. Here, we use a robust shape-from-shading method [15] to estimate surface normals maps (i.e., needlemaps) of the objects in the images. Shape measurements obtained from the estimated needlemaps provide an approximate description of the 3–D geometry of the object's surface. Finally, our results show that the proposed method is able to obtain improvements in classification that go beyond the best rates obtained by individual cues.

The remainder of this paper is organized as follows. In Section 2, we commence by providing a review of the related literature. In Section 3, the details of our cue combination approach are described. Section 4 provides experimental results on two natural image databases. Finally, Section 5 presents our conclusions and plans for future investigation.

2 Related Literature

Recently, there has been considerable developments in part-based classification methods that model the spatial arrangement of object parts [8,7,5]. These methods are inspired by the original ideas proposed by Fischler and Elschlager [10]. For example, Fergus *et al.* [8] proposed a fully-connected part-based probabilistic model for object categorization. The approach is based on the constellation model proposed in [3]. Fergus' approach uses joint probability densities to describe an object's appearance, scale, shape, and occlusion. Shortcomings of the approach include a computational costly parameter estimation as well as a restrictively small number of object parts that can be modeled. The computational cost, in this case, can be addressed by representing the object's spatial structure using tree-structured graphical models [7,6,5].

Yet, most classification methods are based on measurements obtained from a single visual cue. Such an information may not always be available. This limitation can be addressed by combining information from multiple visual cues [13,4,1]. In this paper, we will focus on cue combination methods for object classification. For example, Nilsback and Caputo [13] proposed a cue integration framework based on the linear combination of margin-based classifiers such as vector machines. In another approach, Carbonetto *et al.* [4] combines local image features and region segmentation cues using semi-supervised learning. They also address the problem of selecting reliable local features for classification. However, no spatial structure for the cues is provided (i.e., the visual cues are assumed to be both available and reliable across the entire image). Probabilistic graphical models have also been used for visual cue integration [1], and applied to the problem of direction of figure (DOF) detection and disambiguation. Next, we describe the details of our probabilistic cue combination method.

3 Cue Integration Model

In the description that follows, we drawn our inspiration from recent work on constellation modeling of objects [3] and part-based object classification [6]. In

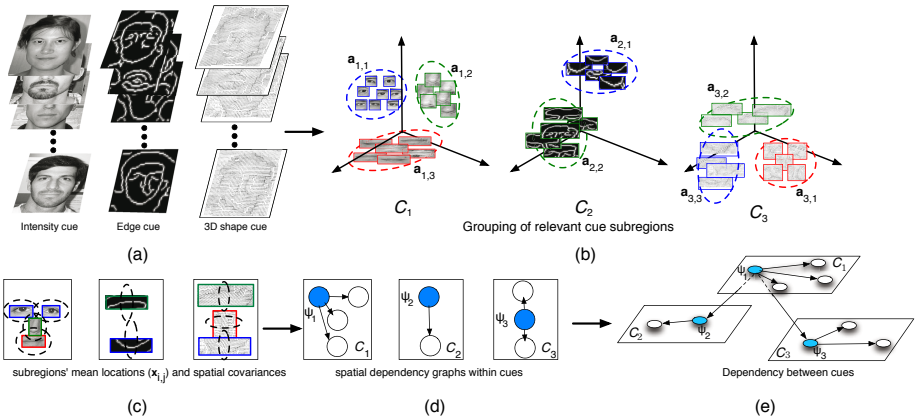

Fig. 1. Visual cue integration

this paper, we show how similar ideas can be applied to the problem of multiple visual cues integration.

We begin by defining the main components of our cue integration model. Let $\boldsymbol{\mathcal{C}} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ represent a set of K visual cues extracted from an image $\boldsymbol{\mathcal{I}}$ of an object (e.g., edge maps, surface normals, color). The goal of our approach is twofold. First, we aim at integrating the information provided by multiple visual cues in a principled manner. Secondly, we will use this integration model for the classification objects in images. Probabilistically, the likelihood of observing a particular image given that an object is at some location can be represented by the distribution $p(\boldsymbol{\mathcal{I}}|\boldsymbol{\mathcal{X}})$ where $\boldsymbol{\mathcal{X}}$ represents a particular spatial configuration of the visual cues associated with the object. From the Bayes' theorem, we obtain:

$$p(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{I}}) \propto \underbrace{p(\boldsymbol{\mathcal{I}}|\boldsymbol{\mathcal{X}})}_{\text{likelihood}} \underbrace{p(\boldsymbol{\mathcal{X}})}_{\text{prior}} \propto \underbrace{p(\boldsymbol{\mathcal{C}}|\boldsymbol{\mathcal{X}})}_{\text{appearance}} \underbrace{p(\boldsymbol{\mathcal{X}})}_{\text{spatial configuration}} \tag{1}$$

In (1), $\boldsymbol{\mathcal{I}}$ was substituted by $\boldsymbol{\mathcal{C}}$ to indicate that the image information will be represented by a set of visual cues. Our cue integration model follows the factorization of Equation 1 suggested by Felzenszwalb and Huttenlocher [7]. The underlying idea in this factorization is that the spatial arrangement of object parts can be encoded into the prior probability distribution while the likelihood distribution encodes the appearance of the object (and its parts). In this paper, we focus ourselves on the representation of both the appearance and the spatial configuration of multiple visual cues.

3.1 Spatial Prior Model

The prior distribution in Equation 1 is described as follows. We assume that each available visual cue $\mathcal{C}_i$ can be subdivided into a number of non-overlapping subregions such that $\mathcal{C}_i = \{(\mathbf{a}_1^{(i)}, \mathbf{x}_1^{(i)}), \ldots, (\mathbf{a}_{N_{\mathcal{C}_i}}^{(i)}, \mathbf{x}_{N_{\mathcal{C}_i}}^{(i)})\}$, where each pair $(\mathbf{a}_j^{(i)}, \mathbf{x}_j^{(i)})$

represents both the local appearance $\mathbf{a}$ and the spatial location $\mathbf{x}$ of the subregion j for the cue $\mathcal{C}_i$. Here, $N_{\mathcal{C}}$ is the total number of subregions within a cue. At this point, we would like to introduce the concept of reliability of a visual cue. Cue reliability can be interpreted as the significance (i.e., contribution) of the cue in the integration process [13]. It can be measured, for example, in terms of the consistency of a cue with respect to the overall recognition rates of the integrated model. In classification problems, it is often the case that some cues are more reliable than others. Moreover, visual cues may not be consistently available and reliable across the entire image. For instance, edge maps may be more reliable at certain image regions while pixel intensity at others.

In this paper, we model the cue integration process and cue reliability using a tree-structured directed acyclic graph [2] in which conditional dependencies represent the implicit cue reliability in the spatial global integration model. Figure 1 illustrates our cue integration concept. Here, the arrows in the graph indicate the conditional dependence between the connected vertices (i.e., less-reliable cues are conditioned to the most reliable one). For simplicity, our model uses the star graph structure suggested by Fergus *et al.* [9]. Here, a particular vertex is assigned to be a landmark vertex $(\mathbf{a}_r^{(i)}, \mathbf{x}_r^{(i)})$. The remaining vertices are conditioned on the landmark vertex. Thus, the dependence between cues in our model is explicitly modeled through the connections between landmark vertices. It should be noted that the dependence between cue regions is based solely on their spatial location as we assume that visual cues are independent with respect to their appearance. We will discuss this assumption in more detail later.

Bayesian Network Factorization. We now introduce the factorization for the prior distribution term in (1). This probability distribution models the spatial interaction among extracted visual cues. When modeling the dependency relationships between cues, we first assume that cues are ordered based on their reliability values. This assumptions allows us to arrange the cues' landmark vertices as a star-shaped tree structure in which landmark vertices of less-reliable cues are conditioned on the landmark vertex of the most reliable cue. For simplicity, we consider the location of the landmark subregion to represent the center of the underlying image cue. Figure 1 illustrates an example of a Bayesian network describing this modeling. Accordingly, the joint distribution for the cues spatial interaction can be derived from the graphical model shown in Figure 1, and is given by:

$$p(\boldsymbol{\mathcal{X}}) = p(\psi_r) \prod_{i \neq r} p(\psi_i | \psi_r) \tag{2}$$

where ψ_i corresponds to spatial configuration of the i-th cue, and the probability distributions that compose Equation 2 are:

$$p(\psi_r) = p(\mathbf{x}_r^{(1)}) \prod_{j \neq r} p(\mathbf{x}_j^{(1)} | \mathbf{x}_r^{(1)}) \tag{3}$$

$$p(\psi_i | \psi_r) = p(\mathbf{x}_r^{(i)}) \prod_{k \neq r} p(\mathbf{x}_k^{(i)} | \mathbf{x}_r^{(i)}) \tag{4}$$

Appearance Model. We now focus ourselves on the appearance term of Equation 1. We assume that the appearance of the visual cues and as well as their corresponding non-overlapping subregions are independent. Under the independence assumption, the appearance likelihood of the combined cues can be factorized into the product of the individual subregions' likelihoods. The likelihood function in (1) becomes:

$$p\left(\boldsymbol{\mathcal{C}}|\boldsymbol{\mathcal{X}}\right) = \prod_{i}^{K} p\left(\mathcal{C}_i|\boldsymbol{\mathcal{X}}\right) = \prod_{i}^{K}\prod_{j}^{N_{\mathcal{C}_i}} p(\mathbf{a}_j^{(i)}|\mathbf{x}_j^{(i)}) \tag{5}$$

Appearance independence is a reasonable assumption for non-overlapping subregions within a single visual cue. However, this independence might be sometimes difficult to achieve. For instance, for visual cues such as color and edges, abrupt changes in color distribution may induce the occurrence of edges on the same image. We plan to study the effects of this assumption in our future work. Next, we describe the learning and recognition stages of our method.

3.2 Learning

The factorization described in Equation 2 and Equation 5 allows for the learning process to be performed in a modular fashion given a set of training images $\{\boldsymbol{\mathcal{I}}_1, \ldots, \boldsymbol{\mathcal{I}}_M\}$. The learning process is divided into two main steps. First, the algorithm estimates the appearance model parameters for each representative subregion in each visual cue layer. Secondly, the parameters representing the spatial configuration of cues are determined. The main learning steps of our algorithm are detailed as follows. For simplicity, we make use of Gaussian densities for conditional probabilities in the model.

Learning the Appearance of Visual Cues' Subregions. In this step, the parameters of the appearance model (Equation 5) are estimated. We commence by extracting a set of visual cues from the training images (e.g., surface normals, edge maps, color, and geometric measurements).

1. **Detect and extract representative subregions in each visual cue.** In this step, each visual cue image is divided into a number of subregions centered at locations provided by an interest feature detector. It should be noted that each visual cue conveys a different type of visual information. Consequently, a cue-specific feature detector should be used for each cue type. For simplicity, we first locate regions of interest in the gray-level image cue using the feature detector described in [12]. We then use these locations to extract multiscale subregions of interest in the remaining visual cues using a Gaussian pyramid approach.
2. **Grouping similar subregions.** The subregions obtained in the previous step are subsequently processed by a semi-supervised learning algorithm to determine the most representative non-overlapping subregions in each cue

map. Here, the method requires two types of input. The first one is a set of positive (i.e., images containing the target object) and a set of negative (i.e., background images) training images. The second input of the method is the number of representative parts to be learned in each cue layer (i.e., $N_{\mathcal{C}_i}$). In our current implementation, we use a modified K-Means clustering method for grouping the representative object parts while giving preference to non-overlapping image subregions. The centroid of the largest K-Means clusters are chosen to be the representative parts of the object for the underlying visual cue. Representative parts of an object are the ones that do not appear frequently in background images.

3. **Appearance likelihood parameter estimation.** Under the Gaussian assumption, the parameters of Equation 5 can be directly estimated by simple calculations of the sample mean vectors and sample covariance matrices of each representative part learned during the clustering step. This step of the learning process is illustrated in Figures 1–(a) and 1–(b).

Learning the Spatial Prior. In this step, our main goal is to estimate the parameters of the spatial prior of the cue integration model (Equation 2). Our method for learning the spatial prior is divided into two main steps. First, for each visual cue, landmark vertices of each cue are chosen to be the ones that can be consistently located in the positive training images dataset. The remaining subregions are conditioned to the landmark ones. The conditioning step is illustrated by the graphs in Figure 1–(d). Secondly, we determine the best global cue dependency configuration using exhaustive search based on the overall classification rate (Figure 1–(e)). We estimate the most likely spatial configuration of the learned parts' locations by selecting the joint probability distribution that allows for the maximum overall recognition rate. The main steps of this stage are described as follows.

1. **Learn the location uncertainty of learned subregions.** The goal of this step is to determine possible locations of the parts previously learned by the algorithm. This is equivalent to a template matching operation. In our implementation, we simply select the image location with maximum likelihood of the part appearance $p(\mathbf{a}_j^{(i)}|\mathbf{x}_j^{(i)})$. The mean location and covariance matrix of each part location is estimated.
2. **Determine the joint Gaussian probability of part locations.** Here, the conditional probabilities of the subregion locations in Equation 2 are estimated using Gaussian joint probability distributions. It can be shown that the conditional distributions relating independent Gaussian distributions are also Gaussian. As a result, the terms $p(\mathbf{x}_k^{(i)}|\mathbf{x}_r^{(i)})$ in (2) take a particularly simple form [2]. Figure 1–(c) shows elliptical shapes illustrates the spatial location uncertainty of each representative part in the model.
3. **Cue probabilities and global model probability.** Once the model parameters for the spatial configuration of subregions within each visual cue

are at hand, the integration of all available visual cues is accomplished by estimating the parameters of Gaussian joint probabilities as described in Equations 3 and 4.

3.3 Recognition and Detection

Once the parameters of the cue integration framework are estimated, the problem of recognizing (and in this case, also locating) an object in an image can then be posed as follows: we seek the location in the image that maximizes the posterior probability of the location of the object given a set of visual cues as given in (1):

$$\boldsymbol{\mathcal{X}}^* = \arg\max_{\boldsymbol{\mathcal{X}}} p(\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{C}}) \tag{6}$$

An exact inference using the model described in (2) is computationally intractable. A possible way to overcome this problem is to make use of approximate inference methods (i.e., belief propagation, expectation-maximization). In this paper, the approach is to first detect every cue individually and then obtain the probability of object configuration using the global relationship among cues. Here, we follow the inference approach suggested in [6]. The recognition stage is accomplished using the following main steps.

1. **Determine part locations.** We begin by extracting a set of image subregions located at interest points in a similar fashion as in the first step of the appearance learning procedure.
2. **Appearance and spatial configuration.** Calculate the appearance likelihood based on the appearance of the parts as described by Equation 5. Determine the probability of spatial configuration of cue parts as in Equation 2. Select the model with the maximum overall posterior probability.

4 Experimental Results

In this section, we assess the potential of our cue combination model for the problem of object recognition. Here, we describe the experimental results performed on two sets of real-world images. The first of these consists of a dataset of marine biofouling organisms (i.e., barnacles). These organisms are usually found attached to the hull of ships and have a dome-like shape. The second dataset used in our experiments consists of images from the Caltech face database. Images samples from both datasets are shown in Figure 2. The choice of the class of images is aimed at demonstrating the ability of our model to integrate both 2–D and 3–D visual cues. The images used in our experiments were divided into subsets of training, validation, and test images. The sizes of the subsets were 100, 100, and 300 images, respectively. Each subset contained an equal amount of object images and background images.

We commenced by processing all images to obtain a set of visual cues represented by maps of pixel intensity, edges, and 3–D shape information. The pixel

Fig. 2. Sample of the images used in our experiments. Row 1: images of barnacles (i.e., marine biofouling organisms). Row 2: face images from the Caltech face database.

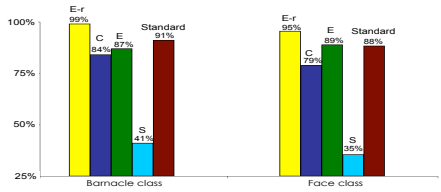

Fig. 3. Recognition rates obtained for barnacles and faces. (E-r) Our probabilistic cue integration model; (C) Classification rate using gray-level intensity only; (E) Classification rate using edge-map only; (S) Classification rate using estimated surface normals only; (Standard) Simple linear combination of cues.

intensity map consisted of simple gray level versions of the images. The edge map information was obtained using the Canny edge detector. Finally, the third visual cue was obtained from surface normals (needlemaps) estimated using the robust shape-from-shading method proposed by Worthington and Hancock [15]. Surface normals from shape-from-shading as a single visual cue have been recently used for object recognition [15].

In our experiments, we created two sets of graphical models. For the barnacle class, we created two, four, and one part star-graph models to represent the gray-level intensity, edge, and 3–D shape cues, respectively. The choice of spatial configuration was determined automatically to maximize the classification rates for each individual visual cue.

We also experimented our method on images from the Caltech face database. For these images, the edge cue was the most reliable one. For the face class, we used two-vertex graphs for both the gray-level pixel intensity and the edge map cues. The 3–D information was represented by a single landmark vertex graph.

Finally, our experiments include a comparison between our cue integration framework and the one described in [11]. In [11], the final recognition score is calculated as the linear combination of the individual cues' maximum posterior probabilities. Additionally, we provide a comparison between

the recognition results obtained using our model and the results obtained using each cue individually. Figure 3 shows the results of this comparative study. The results indicate that the combined model provides significantly better recognition rate when compared to results obtained based on single cues only. Our preliminary results also show that our method outperforms the cue integration method based solely on linear combination for the class of images used in this paper.

5 Conclusions and Remarks

In this paper, we proposed a probabilistic model for the integration of visual cues for object recognition. We drew our motivation from recent probabilistic part-based models for object recognition. Here, we derived a Bayesian framework for multiple cue integration. Our model was able to represent the different contributions of each cue in the recognition process at different spatial locations in the object. We also combined information from 2–D and 3–D modalities. Finally, our experiments showed the effectiveness of our method for object recognition.

The work presented in this paper represents an attempt to accomplish cue integration in a principled way. The use of a probabilistic framework that describes cue dependencies consists of a natural integration approach that allows for the inclusion of prior information while permitting the learning of models from training data.

The proposed method have many avenues for improvement. The experiments presented in this paper are preliminary. However, they serve to show the potential of the proposed cue integration method. The use of only three cues is limiting and a larger number of cues should be added to assess the behavior of the method. A study of the selection of difference dependences and cue would also be helpful. Additionally, the use of only two object classes is somehow insufficient. We plan to extend the experiments to a larger number of object classes.

The use of 3-D shape information is clearly a desirable feature in a cue integration framework. In this paper, we added shape information in the form of surface normals provided by a shape-from-shading algorithm. Unfortunately, using surface normals directly does not provide an appropriate treatment of the 3-D information due to often low quality of the measured data. Improvements can probably be obtained by using alternative shape representations such as shape-index and spin-images.

Finally, we have arbitrarily selected the types of cues to use in our model. In principle, any visual cue can be used by the method. However, for simplicity, we have selected a set of commonly used cues in image analysis. A more comprehensive study of effective cues and their representation within the cue integration framework is needed. Work addressing the above issues is currently underway and it will be published in due course.

References

1. Baek, K., Sajda, P.: A probabilistic network model for integrating visual cues and inferring intermediate-level representations. In: IEEE Workshop on Statistical and Computational Theories of Vision, Nice, France, IEEE Computer Society Press, Los Alamitos (2003)
2. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics), Secaucus, NJ, USA. Springer, New York (2006)
3. Burl, M.C., Weber, M., Perona, P.: A probabilistic approach to object recognition using local photometry and global geometry. In: European Conference on Computer Vision, London, UK, vol. II, pp. 628–641 (1998)
4. Carbonetto, P., Dorkó, G., Schmid, C., Kück, H., de Freitas, N.: A semi-supervised learning approach to object recognition with spatial integration of local features and sementation cues. In: Towards category-level object recognition, Springer, Heidelberg (2006)
5. Carneiro, G., Lowe, D.: Sparse flexible models of local features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3953, pp. 29–43. Springer, Heidelberg (2006)
6. Crandall, D., Felzenszwalb, P.F., Huttenlocher, D.P.: Object recognition by combining appearance and geometry. In: Toward Category-Level Object Recognition, pp. 462–482 (2006)
7. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. Int. J. Comput. Vision 61(1), 55–79 (2005)
8. Fergus, R., Perona, P., Zisserman, A.: Weakly supervised scale-invariant learning of models for visual recognition. Int. J. Comput. Vision 71(3), 273–303 (2007)
9. Fergus, R., Perona, P., Zisserman, A.: A sparse object category model for efficient learning and exhaustive recognition. In: CVPR 2005. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (June 2005)
10. Fischler, M., Elschlager, R.: The representation and matching of pictorial structures. IEEE Transactions - Computers 22, 67–92 (1977)
11. Hayman, E., Eklundh, J.-O.: Probabilistic and voting approaches to cue integration for figure-ground segmentation. In: European Conference on Computer Vision-Part III, London, UK, pp. 469–486 (2002)
12. Kadir, T., Zisserman, A., Brady, M.: An affine invariant salient region detector. In: European Conference on Computer Vision, vol. I, pp. 228–241 (2004)
13. Nilsback, M., Caputo, B.: Cue integration through discriminative accumulation. In: Conf. on Computer Vision and Pattern Recognition, pp. II: 578–585 (2004)
14. Tanaka, K., Saito, H., Fukada, Y., Moriya, M.: Coding visual images of objects in the inferotemporal cortex of the macaque monkey. Journals of Neurophysiology 66(1), 170–189 (1991)
15. Worthington, P.L., Hancock, E.R.: Object recognition using shape-from-shading. IEEE Trans. Pattern Anal. Mach. Intell. 23(5), 535–542 (2001)