Transinformation for Active Object Recognition

Bernt Schiele
MIT – The Media Laboratory
Cambridge, MA 02139, USA
bernt@media.mit.edu

James L. Crowley
GRAVIR, Institute IMAG
38031 Grenoble, France
jlc@imag.fr

Abstract

This article develops an analogy between object recognition and the transmission of information through a channel based on the statistical representation of the appearances of 3D objects. This analogy provides a means to quantitatively evaluate the contribution of individual receptive field vectors, and to predict the performance of the object recognition process. Transinformation also provides a quantitative measure of the discrimination provided by each viewpoint, thus permitting the determination of the most discriminant viewpoints. As an application, the article develops an active object recognition algorithm which is able to resolve ambiguities inherent in a single–view recognition algorithm.

1 Introduction

We have generalized the color histogram approach of Swain and Ballard [10] to histograms over vectors of local neighborhood operators, which we call multidimensional receptive field histograms [7]. A first recognition algorithm is given by histogram matching providing fast and robust recognition of objects in the presence of image plane rotations, scale changes and viewpoint changes. [8] develops a probabilistic object recognition algorithm which recognizes objects from a small portion of the object. Experimental results show that the algorithm recognizes reliably multiple objects in cluttered scenes [7] from a database of 100 objects.

Section 2 develops a general framework for the statistical representation of objects by multidimensional receptive field histograms. The underlying idea is that multidimensional receptive field histograms represent reliably a certain appearance of an object whereas a collection of histograms approximates the probability density function of the 3D object. A probabilistic object recognition as a single view recognition algorithm is described. This algorithm recognizes objects from a small visible portion of an object. The algorithm is therefore robust to background changes and partial occlusion enabling multiple object recognition in cluttered scenes (see [7] for experiments).

Section 3 expresses object recognition in terms of the transmission of information through a channel. In this context we obtain a quantitative measurement for the overall uncertainty of recognition, the average "noise" and the overall quality of recognition. In particular, we emphasize the *transinformation* of object recognition, which can be employed for the quantitative evaluation of the chosen measurement set.

Section 4 shows an application of the developed analogy between object recognition and information theory: using transinformation of single viewpoints we identify the most salient viewpoints of an object. Section 5 uses such salient viewpoints for the definition of an active object recognition algorithm which moves the camera to the most discriminant viewpoint of a hypothesized object in order to verify the presence of the hypothesized object. Experiments on the Columbia image database (100 objects from 72 viewpoints [4]) show that the proposed algorithm is able to resolve ambiguities which are inherent in a single–view recognition algorithm. By integrating multiple views and in particular the most discriminant views of objects, the algorithm uses not only 2D information but also 3D information of objects. By introducing several verification steps a nearly 100% recognition rate is obtained.

2 Statistical object representation

This section motivates the use of multidimensional receptive field histograms for the statistical representation of the appearances of objects. The main idea is to represent 3D objects by the probability density function of 2D local characteristics, which can be calculated reliably from images of the objects. In this article we use Gaussian derivatives as local characteristics, but the same method can and should be applied to other local descriptors. The transinformation of object recognition (section 3) can be used for the evaluation of the employed local descriptors.

Using a fixed measurement set M of local characteristics m_k, the probability density function over the measurement set M for a certain object o_n varies with the changes of the appearance of the object. Pos-

sible changes include: arbitrary *3D rotations* R and *translations* T of the object, *partial occlusions* P, *light changes* L (including changes in intensity, color and direction of the light) and *noise* N (different types of signal disturbances can be modeled as "noise"). By writing the probability density of the object o_n, parameterized by these variables, we obtain:

$$p(M|o_n, R, T, P, L, N) \quad (1)$$

Since it is not very attractive (and in general difficult) to estimate the "complete" probability density function, we want to reduce the number of free parameters of the probability density function. The most efficient way is to choose local characteristics which are invariant with respect to different parameters. Such invariant features are used by many researchers [3] and applied successfully in various ways. Unfortunately the obtained invariants are very restrictive to certain types of objects. Local characteristics which are robust with respect to certain changes also reduce the number of free parameters. Robust means that local characteristics change slowly with certain changes, implying quasi–invariance. The advantage is that many local characteristic can be calculated in a robust manner without being invariant in general.

With respect to changes of *light* and *noise* we apply local characteristics which are robust to such changes. The analysis of the robustness of the employed local characteristics to such changes is very important [7].

It is difficult to model *partial occlusion* in a general way. In section 2.1 we propose a probabilistic object recognition approach which recognizes objects by the observation of only a small portion of the object. This makes the recognition robust to partial occlusion.

Three degrees of freedom are given by the *translation* vector T of the object. In order to avoid the difficult and time consuming correspondence problem we do not represent the 2D position of the local measurements in the image plane. This implies that we do not need to calculate correspondence as well as it reduces the dimensionality of the probability density function by two dimensions. This makes the estimation of the probability density function feasible due to the amount of training samples which is provided by images of an object. A typical 512×512 image of an object provides $500^2 = 250,000$ training samples.

The remaining component of the translation vector is chosen perpendicular to the image plane and is directly related to the size of the object in the image. We use directly the size (or scale) σ of the object in the image as representation of this component of the translation vector. If we want to move the camera to a particular position in space (relative to the object) we have to know (approximately) the relation between the size σ of the object and the distance of the object to the camera (which implies an approximate calibration of the camera). In this article we consider that the relation is known (i.e. can be estimated off-line during the learning phase of the probability density function, where we assume that the distance between the object and the camera is known).

An arbitrary *rotation* of an object can be represented by three degrees of freedom of the probability density function. If we do not want to restrict the applicability of the approach to certain object classes (with possible self–occlusion, free-form objects) we have to represent at least two degrees of the rotations of an object. We can use local characteristics which are invariant to rotation perpendicular to the image plane [9] (by loosing rotational information about the object). But no local characteristics exist which are invariant to arbitrary 3D rotations. Therefore, we have to consider at least two, in general all three components of the rotation.

What remains from the probability density function 1 are three (or two) components of the rotation R and one component of the translation (represented by the size σ of the object), called the pose $S = (\sigma, R)$:

$$p(M|o_n, S) \quad (2)$$

Different possibilities exist for the representation of this density function. Hornegger and Niemann [2] use parameterized mixtures of multivariate Gaussian distributions including a feature transform. This modeling has been shown to be appropriate for point features but cannot be assumed for more general local characteristics. Another possibility is to use kernel functions [5], which typically allow to generalize from training samples, without representing the training samples well. Histogramming, which we have chosen for the presentation of the density function, represents the training samples very well. In the context of histogramming, generalization capability can be obtained by a sufficient number of training samples. In our case, the number of training samples is directly given by the size of the object in the image. As a result, the order of training samples – respectively image measurements – is in the same order as the size of the object. Histogramming is therefore expected to generalize from the training samples.

Consequently, we have chosen to represent the density function by multidimensional histograms over the measurement set M, where each histogram corresponds to a particular rotation R and to a certain scale of the object σ. A histogram of a particular view is defined by: $H(M|o_n, S_j)$, with $S_j = (\sigma_j, \alpha_j, \beta_j, \gamma_j)^T$ fixed for a particular histogram. σ_j represents the scale of the object and $(\alpha_j, \beta_j, \gamma_j)$ an arbitrary 3D rotation. The representation of an object from any view is given by a collection of several histograms distributed over all possible views. In order to reduce

the number of images per object we employ two properties of Gaussian derivatives, namely the steerability [1, 7] to image plane rotation and the equivariance property to scale. As a result we calculate histograms corresponding to arbitrary scales and arbitrary image plane rotations from a single image of an object. Using these properties of Gaussian derivatives we have to consider only two degrees of rotations (corresponding to a viewing sphere with constant radius).

2.1 Probabilistic object recognition

In the context of probabilistic object recognition we are interested in the calculation of the probability of the object o_n and its pose S_j given a local image region Reg: $p(o_n, S_j|Reg) = p(o_n, S_j|\bigwedge_k m_k)$, with m_k local image measurements inside Reg. In [8] we developed the following formula:

$$p(o_n, S_j|\bigwedge_k m_k) = \frac{\prod_k p(m_k|o_n, S_j)}{\sum_{n',j'} \prod_k p(m_k|o_{n'}, S_{j'})} \quad (3)$$

The probabilities $p(m_k|o_n, S_j)$ are directly given by the multidimensional receptive field histograms. Therefore, equation 3 shows a calculation of the probability for each object o_n only based on the multidimensional receptive field histograms of the N objects.

It is important to note that the locations of the measurements can be chosen arbitrarily. Therefore, the technique is fast (only a certain number of local receptive field vectors have to be calculated) and robust to occlusion (the approach is strictly local). Furthermore the technique works without correspondence between the object database and the test image.

Experiments [8] show that a visible object portion of 40% is sufficient for the recognition of a database of 103 objects. The algorithm is therefore robust with respect to partial occlusion. In the reminder of the article we hypothesize an object $\hat{o_n}$ and its pose $\hat{S_j}$ from a single view. This is done by maximizing the probability $p(o_n, S_j|Reg)$ for an arbitrary region Reg:

$$(\hat{o_n}, \hat{S_j}) : \max_{n,j} p(o_n, S_j|Reg)) \quad (4)$$

3 Application of information theory to object recognition

In the following we express the object recognition process in terms of the transmission of information through a (noisy) channel. Even though this analogy is most appropriate for a statistical object representation it may be applied to a wide variety of object recognition processes.

Section 3.1 summarizes basic concepts from information theory [6]. Section 3.2 discusses the interpretation of different entropies in the context of object recognition. Section 3.3 introduces the concept of transinformation of object recognition, which can be applied for the evaluation of the employed measurement set. Experiments show that we can predict the performance of the object recognition process.

3.1 Information measurement

The sample space Ω_X with its mutually exclusive events x_n forms a *complete finite probability scheme* if it is true that: $\bigcup_{n=1}^{N} x_n = \Omega_X$ and $\sum_{n=1}^{N} p(x_n) = 1$. The fundamental problem of interest in information theory is to define a measure of *uncertainty* for such a probability scheme. Shannon and Wiener have suggested to use the well known equation:

$$H(X) = -\sum_{n=1}^{N} p(x_n) \log(p(x_n)) \quad (5)$$

Using the quantity $I(x_n) = -\log(p(x_n))$ as the measure of *self–information* of the event x_n we can interpret $H(X)$ as the average self-information of each event x_n. The object set $\Omega_O = \bigcup_{n=1}^{N} o_n$ (as well as the measurement set Ω_M) form a complete finite probability scheme. Therefore, we can define the average self–information $H(O)$ of each object o_n: $H(O)$ (as well as $H(M)$).

In the context of the transmission of information through a channel we need to know the relation between the input symbols and the output symbols. The measurement $H(X)$ of *uncertainty* or *information* is therefore generalized to a 2D discrete finite probability scheme. Such a probability scheme is given by two sample spaces Ω_X and Ω_Y where complete event sets x_n and y_k are selected. Each event x_n of Ω_X may occur in conjunction with any event y_k of Ω_Y. Therefore, the product space $\Omega_X \times \Omega_Y$ forms a complete set of events with probability matrix $P(X \wedge Y)$.

Therefore, we have three complete probability schemes, namely $P(X)$, $P(Y)$ and $P(X \wedge Y)$. We can define three corresponding entropies: $H(X)$, $H(Y)$, and the joint entropy $H(X \wedge Y)$. We can also define the conditional entropies $H(X|Y)$ and $H(Y|X)$ [6].

3.2 Application of information theory to object recognition

The previous section introduced five entropies for a 2D discrete finite probability scheme without giving their interpretation. In information theory a 2D probability scheme is used to describe a communication network: x_n are the possible inputs (or symbols of the input alphabet) and y_k are the possible outputs of the network. Each input x_n is "transformed" by the communication channel to possible outputs y_k, whereas the joint probability $P(X \wedge Y)$ describes the characteristics of the channel.

In the context of object recognition the possible "inputs" are the different objects o_n and the possible "outputs" are the measurements or symbols m_k which we extract from the image of an object. The channel corresponds to the transformation of the objects to

the measurement space. Therefore, the communication channel corresponds to the recognition process as a whole. The characteristics of the recognition process are given by the joint probability. Using this analogy between a communication network and the object recognition process we can interpret the five entropies in the following way:

- $H(O)$: average information of each object o_n,
- $H(M)$: average information of each feature m_k,
- $H(O \wedge M)$: overall uncertainty of the recognition,
- $H(M|O)$: indication of the average "noise" or error of the recognition process,
- $H(O|M)$: indication of overall quality of the recognition process. The smaller $H(O|M)$ the better the object set O can be recognized with the measurement set M.

Since we consider fixed a priori probabilities $p(o_n)$, the entropy $H(O)$ is constant. The other entropies change with respect to the measurement set M.

3.3 Transinformation of recognition

In information theory the mutual information contained in the event pair (x_n, y_k) is the basis to calculate the *transinformation* of the channel. By applying the analogy between the communication network and the object recognition process we can calculate the average transinformation of all object/measurement pairs (o_n, m_k):

$$T(O,M) = \sum_{n,k=1,1}^{N,K} p(o_n \wedge m_k) \log \frac{p(o_n \wedge m_k)}{p(o_n)p(m_k)} \qquad (6)$$

This entropy indicates a measure of the information transmitted through the channel (= recognition process). For this reason it is known as *transinformation* of the channel. We can easily show that:

$$T(O,M) = H(O) - H(O|M) \qquad (7)$$

$$= H(M) - H(M|O) \qquad (8)$$

Following equation 7, the information transmitted by the channel (respectively by the recognition process) can be maximized by the minimization of $H(O|M)$ (assuming $H(O)$ to be constant). Or by using equation 8 we can say, that the maximization of $H(M)$ and the minimization of $H(M|O)$ result in the maximization of the transinformation.

The most interesting idea of using transinformation is the possibility to compare different measurement sets M for an object set O (or for any subset of O).

In order to illustrate the application of the transinformation for the evaluation of different measurement sets we have calculated the transinformation for 100 objects as a function of different filter–combinations at different resolutions. In figure 1 three different measurement combination are used: *Dx-Dy* (first Gaussian derivatives in $x-$ and in $y-$directions), *Mag-Lap* (magnitude of the first derivative and Laplace operator) and *Dx-Dy-Lap* (first derivatives in $x-$ and $y-$directions and Laplace operator). We use a standard definition of Gaussian derivatives [8]. For each of the filter combinations we calculated histograms with different resolutions between 2 and 128 cells per histogram axis.

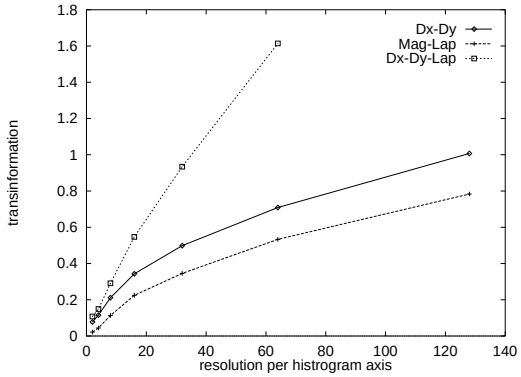

Figure 1: The transinformation for 100 objects and for different measurement sets. The horizontal axis shows different resolutions for the histogram axis and the vertical axis show the transinformation for a particular resolution

Figure 1 shows the strong influence of the measurement sets onto the transinformation. A significant increase of the transinformation is obtained for the 3D measurement set *Dx-Dy-Lap* relative to the 2D measurement sets *Dx-Dy* and *Mag-Lap*. This can be explained by the fact that *Dx-Dy-Lap* contains one more independent dimension. This increase is even more important by adding more independent dimensions. It is interesting that the results of *Dx-Dy* and *Mag-Lap* are qualitatively similar. Nevertheless, *Dx-Dy* gives better results than *Mag-Lap* since it preserves the rotational information.

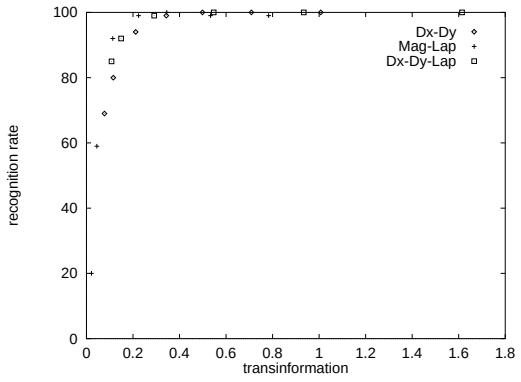

Figure 2: The relation between the transinformation and the recognition of 100 objects

Figure 2 shows the relation between the transinformation and the recognition rate. We used the same measurement sets and the same database images as in figure 1. We used 100 test images with a viewpoint change of 5°. The figure shows the strong relation between the transinformation and the recognition rate. We are therefore convinced that we can predict the quality of recognition based on the transinformation of object recognition.

4 Choice of salient viewpoints

In the previous section we showed the application of information theoretical concepts to the process of object recognition. In this section we apply this analogy in the context of viewpoint planning for object recognition. More precisely we calculate the transinformation of each viewpoint of an object in order to choose the most "significant" viewpoints of an object. Section 5 uses these salient viewpoints in an active object recognition approach.

4.1 Transinformation of a single viewpoint

Section 3.3 defined the transinformation of the object recognition process based on the event pairs (o_n, m_k). In the following we develop an equation for the choice of the most significant viewpoint(s) of an object. Therefore we rewrite equation 6:

$$T(O,M) = \sum_{n=1}^{N} p(o_n) \sum_{k=1}^{K} p(m_k|o_n) \log \frac{p(m_k|o_n)}{p(m_k)}$$

The transinformation can be interpreted as the average transinformation for an object o_n which we define as: $T(o_n, M) = \sum_{k=1}^{K} p(m_k|o_n) \log \frac{p(m_k|o_n)}{p(m_k)}$

The probability $p(m_k|o_n)$ corresponds to the "complete" probability density function. Since we want to calculate the transinformation of an individual viewpoint of an object with respect to all other viewpoints of all other objects, we define the transinformation of an object o_n at a certain pose S_j as follows:

$$T(o_n, S_j, M) = \sum_{k=1}^{K} p(m_k|o_n, S_j) \log \frac{p(m_k|o_n, S_j)}{p(m_k)} \quad (9)$$

The most significant viewpoints of an object o_n are then given by the maxima (over j) of equation 9.

5 Active object recognition

The principal idea of our active recognition algorithm is to hypothesize the object identity and its pose from a test-image (single view recognition). Based on this estimation the camera is moved to the most discriminant viewpoint(s) of the hypothesized object. The information gathered from the new viewing direction is then used for verification of the hypothesis.

Experimental results (section 5.2) show that the algorithm can resolve all (except one) of the ambiguities which are inherent to single–view recognition. Most interestingly the algorithm incorporates 3D information from the statistical representation of the object. Remaining errors are therefore 3D consistent.

5.1 Active object recognition algorithm

The active recognition algorithm contains 3 steps:

Hypothesis generation: an object and pose hypothesis is generated by calculating the probabilities $p(o_n, S_j|Reg)$ for a local image region Reg of the test image. The hypothesized object $\hat{o_n}$ with its pose parameters $\hat{S_j}$ are given by equation 4.

Camera movement: the camera is moved to the salient viewpoint of the object $\hat{o_n}$ which is calculated off-line by equation 9. The movement of the camera is calculated on the basis of the difference ΔS between $\hat{S_j}$ and the pose of the salient viewpoint.

Verification of the hypothesis: using equation 4 a new object and pose hypothesis is generated. For verification we can use two conditions. Firstly, the hypothesized object should be the same before (time $(t-1)$) and after (time (t)) the camera movement:

$$\hat{o_n}(t) = \hat{o_n}(t-1) \quad (10)$$

Additionally, we can use ΔS in order to obtain a prediction of the new pose estimate:

$$\hat{S_j}(t) = \hat{S_j}(t-1) + \Delta S(t-1) \pm \epsilon(S) \quad (11)$$

where $\epsilon(S)$ corresponds to the allowed error. In order to obtain recognition with minimal false positives and minimal false negatives at the same time, we can use several verification steps.

The proposed active object recognition algorithm moves the camera to the most discriminant viewpoint of the hypothesized object. By doing so we can expect to verify the hypothesized object when it is the correct object. On the other hand, we expect to reject the hypothesized object whenever it is not the correct one. Ambiguities in the database, which cannot be solved by a single–view recognition system are expected to be solvable by the proposed approach, since the salient viewpoints are selected relative to the object database.

An attractive property of the proposed active object recognition process is that not only 2D information (a single image) but also 3D information of the statistical object representation is used for verification. That implies that errors of the proposed approach should be only possible between objects which are 3D–compatible.

5.2 Experimental results

The described experiment is based on the Columbia image database of 100 objects [4], containing 72 viewpoints for each object (the database contains color images which we converted to grey scale images). Unfortunately the database contains only one rotational freedom (which we will call α in the following) for the objects. That means that three parameter of the pose estimation (namely β, γ and σ) are not considered

here. The reason for using the database was that we can simulate camera movements by "turning" the object in front of the camera. Therefore, we can validate the proposed algorithm without depending on uncertainties of camera movement and calibration (in order to move the camera relative to the object we have to calibrate the camera at least approximately).

Half of the images (100×36) are used as database. The remaining half are used as test images. For each of the database images we calculate a 4D histogram of the first Gaussian derivatives in x– and in y–direction at two different scales (namely for $\sigma = 2.0$ and $\sigma = 4.0$). In this particular experiment we used a resolution of 32 cells per histogram axis. It is interesting to note that even though the theoretical number of histogram cells is 32^4, in average only 2300 nonzero cells have to be represented per histogram. For each object we have calculated the most and the second most discriminant viewpoint.

verification steps	recognition	number of errors
0	98.22 %	64
1	99.08 %	33
2	99.97 %	1

Table 1: Experimental results on the Columbia images database of 100 3D-objects. For comments see text.

Table 1 shows results which we obtained by the application of the active object recognition algorithm introduced above. In order to make the recognition task more difficult, we choose an image region of 40% of a test image as local region *Reg* for the calculation of the probabilities $p(o_n, S_j|Reg)$. Without using any verification (first line of the table) we obtain 64 misclassifications of the 3600 test images. By using one verification step the number of misclassifications is reduced to 33. By using two verification steps we can resolve all ambiguities but one. These results validate the applicability of the proposed active recognition approach. (During the verification steps we allowed an error $\epsilon(\alpha)$ of up to 10°).

Figure 3: The 5 objects which are confused by the recognition algorithm with one verifications step

It is interesting to note that all misclassifications produced with one verification step are 3D consistent. Figure 3 shows the five objects which have been confounded. Most errors (23 of 33) are made between the last two objects (which can be distinguished in the original database by their color). All five objects are 3D consistent because they are all cuboids. We may discriminate these objects by using color information or other appropriate measurement vectors (which can be evaluated by the transinformation introduced in section 3.3). The only misclassification obtained with two verification steps is between the last two object of figure 3.

6 Conclusion

This article motivates the statistical representation of 3D objects by a collection of multidimensional receptive field histograms. Each histogram represents an appearance of the object. Based on such a statistical representation, object recognition is expressed as the transmission of information through a communication channel. This analogy permits to apply several concepts of information theory to object recognition: the average information of each local characteristic, the overall uncertainty of the recognition process and the average error of the recognition process.

Based on the analogy between the transmission of information through a channel and object recognition we can calculate the *transinformation* of the recognition process. Experiments show the applicability of transinformation for the quantitative evaluation of measurement (or feature) sets. A second application of the transinformation is the determination of the most discriminant viewpoints of an object. Based on such viewpoints the article defines an active recognition algorithm which is able to resolve ambiguities which are inherent in a single–view recognition approach. This algorithm incorporates 3D information of the objects entirely based on 2D image measurements of the object.

References

[1] W.T. Freeman and E.H. Adelson. The design and use of steerable filters. *PAMI*, 13(9):891–906, 1991.

[2] J. Hornegger and H. Niemann. Statistical learning, localization and identification of objects. In *ICCV*, 1995.

[3] J.L. Mundy and A. Zissermann, editors. *Geometric Invariance in Computer Vision*. MIT Press, 1992.

[4] S.A. Nene, S.K. Nayar, and H. Murase. Columbia object image library (coil-100). Technical Report CUCS-006-96, Columbia University, 1996.

[5] K. Popat and R.W. Picard. Cluster–based probability model applied to image restoration and compression. In *ICASSP*, 1994.

[6] F.M. Reza. *An Introduction to Information Teory*. Dover Publications, New York, 1994.

[7] B. Schiele. *Object recognition using multidimensional receptive field histograms*. PhD thesis, I.N.P. Grenoble, 1997. English translation.

[8] B. Schiele and J.L. Crowley. Probabilistic object recognition using multidimensional receptive field histograms. In *ICPR'96, Vol.B*, pages 50–54, 1996.

[9] C. Schmid and R. Mohr. Combining grayvalue invariants with local constraints for object recognition. In *CVPR*, 96.

[10] M.J. Swain and D.H. Ballard. Color indexing. *IJCV*, 7(1):11–32, 1991.