Transinformation of Object Recognition and its Application to Viewpoint Planning

Bernt Schiele [1] and James L. Crowley

GRAVIR, Institute IMAG, 46, av. Félix Viallet, 38031 Grenoble, FRANCE

This article develops an analogy between object recognition and the transmissions of information through a channel. This analogy is based on the statistical representation of the appearance of 3D objects by several multi-dimensional receptive field histograms. The analogy between transmission of information and object recognition provides a means to quantitatively evaluate the contribution of individual receptive field functions, and to predict the performance of the object recognition process using receptive field histograms. 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. The algorithm incorporates 3D information of an objects appearance entirely based on 2D measurements in images of the object.

Key words: Statistical object recognition; Active vision; View point planning; Information theory.

1 Introduction

This article has three parts: section 2 motivates the use of multidimensional receptive field histograms for the statistical representation of objects. The underlying idea is that multidimensional receptive field histograms represent reliably a certain appearance of an 3D object whereas a collection of several histograms represents an approximation of the probability density function of the 3D object. At the end of section 2 we will describe a single view recognition algorithm based on multidimensional receptive field histograms.

[1] email: Bernt.Schiele@imag.fr

In section 3 we express the object recognition process in terms of the transmission of information through a noisy channel. This analogy is most appropriate in the context of statistical representation of objects but can be applied to a wide variety of object recognition processes. In this context one obtains a quantitative measurement for the overall uncertainty of the recognition process, the average "noise" and the overall quality of the recognition process. Furthermore one can define the capacity, the redundancy and the efficiency of the object recognition process. In particular we want to emphasize the *transinformation* of the object recognition process. It can be employed for the quantitative evaluation of the chosen measurement set as well as for the determination of the most discriminant viewpoints of an object (section 4).

Section 5 finally shows an application of the developed analogy between object recognition and information theory: based on the most discriminant viewpoints (section 4) we define 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 nearly all 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. Therefore the remaining errors from the active recognition algorithm are 3D consistent. By introducing several verification steps one incorporates more 3D information so that a nearly 100% recognition rate is obtained.

2 Statistical object representation

This section introduces and motivates the use of multidimensional receptive field histograms for statistical representation of the appearance 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 (for example Gabor filters, grey value moments or geometric invariants). In section 3 we will emphasize the transinformation of the object recognition process as a quantitative measurement for the evaluation of the employed local descriptors (with respect to the considered objects).

In the following we develop a statistical object representation based on the probability density function of the local characteristics of an object o_n. Let's consider we have chosen 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 will vary with the changes of the appearance of the object. Therefore we have to model the possible changes of the objects appearance within the probability density function. Possible changes include:

Rotation: arbitrary 3D rotation of the object in space. An arbitrary rotation can be described by three parameters.

Translation: three parameter are necessary in order to describe an arbitrary translation of an object.

Partial occlusion: it is more difficult to identify the number of parameter (or degrees of freedom) for partial occlusion: firstly an object can be occluded by another known or unknown object. Secondly the portion and the location of the occlusion can vary. And thirdly we would like to model a scene of multiple objects as a special case of partial occlusion.

Light changes: light changes can concern the light intensity, light color, light direction and also the number of light sources.

Noise: different types of disturbances of the signal can be modeled as some type of "noise". Examples include quantization error, signal noise and blur.

By writing the probability density function of the object o_n, parameterized by these variables, one obtains:

$$p(M|o_n, R, T, P, L, N) \tag{1}$$

where M is the set local characteristic m_k, o_n is the label of the object (or object class), R describes the rotation of the object, T the translation, P the condition of the partial occlusion, L the light conditions and N the noise.

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. Probably the most tempting way to reduce the number of free parameters is to choose local characteristics which are invariant in relation to the 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. Furthermore these invariants usually need the calculation of high order derivatives which are in general relatively unstable. Another disadvantage is often that the invariants are global which makes it difficult to deal with partial occlusion.

Another way to reduce the number of free parameters of the probability density function 1 is to use local characteristics which are robust in relation to certain changes. Robust means that the local characteristics are changing slowly with the considered changes. Typically one can identify a certain range of changes where the local characteristic is nearly constant. If one can identify or assume

this range of changes, one can think of these robust local characteristics as nearly-invariant descriptor. We are convinced that many local characteristic can be calculated in a robust manner without being invariant in general.

In relation to changes of *light* and *noise* we will apply local characteristics which are robust (in the sense described above) to such changes. Therefore the analysis of the robustness of the employed local characteristics to such changes is very important.

As mentioned above, it is relatively difficult to model *partial occlusion* in a general way. Hornegger and Niemann [2] propose to model partial occlusion as a particular object: the background. By introducing a probability for the background – which is related to the observed portion of the object – the probability of the presence of an object in the image can be calculated. The recognition process therefore estimates not only the object's label and its pose but also the portion of occlusion, which makes the recognition process elegant but relatively time consuming. In [8] we have proposed a probabilistic object recognition approach which is able to recognize objects by the observation of only a small portion of the object. This makes the recognition process robust to partial occlusion. As a result we do not have to consider partial occlusion in the modeling of the probability density function of an object. Furthermore the recognition process is relatively fast: the recognition time for a database of 100 objects is less than one second on a usual workstation without special hardware.

Three degrees of freedom are given by the *translation* vector T of the object. In the following we will assume, that the first two components of the vector are chosen to be parallel to the image plane. The third component is chosen to be perpendicular to the image plane and is related to the size (or scale) of the object in the image. For the representation of objects by its probability density function we will omit the first two components which are parallel to the image plane. Two reasons justify this choice: first of all we obtain high recognition rates by doing so. The second reason is the following: since we do not represent the translational information in the probability density function we do not have to solve the correspondence problem between a database object and a test object, which is usually difficult and time consuming.

The third component of the translation vector can be chosen to be perpendicular to the image plane. Therefore this component is directly related to the size of the object in the image. We use directly the size (or scale) s of the object in the image as representation of this component of the translation vector. If one wants to move the camera to a particular position in space (relative to the object) one has to know (at least approximately) the relation between the size s 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 for the probability density function, where we assume to know the distance of the object to the camera).

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

What remains from the original probability density function 1 are three (or two) components of the rotation and one component of the translation (represented by the size s of the object):

$$p(M|o_n, R, s) \tag{2}$$

Different possibilities exist in order to represent this density function. Hornegger and Niemann [2] use parameterized mixtures of multivariate Gaussian distributions including a feature transform. These statistical models consider the statistical behavior of features, feature matching, as well as the projection from the model into the image space. This modeling has been shown to be appropriate for point features but cannot be assumed for more general local characteristics. We have chosen to represent the density function by multi-dimensional histograms over the measurement set M, where each histogram corresponds to a particular rotation R and to a certain scale of the object s. In order to obtain the correct histogram of an arbitrary rotation of the object we have to take several images of the object. But in order to reduce the number of images we employed in [7,8] the steerability of the Gaussian derivatives [1] to image plane rotation and scale. Using this property of Gaussian derivatives one has to consider only two degrees of rotations. A histogram of a particular view is then defined by:

$$H(M|o_n, r_i, s_m) \tag{3}$$

where $r_i = (\alpha_i, \beta_i, \gamma_i)^T$ and s_m are fixed for a particular histogram. r_i represents the three degrees of freedom of an arbitrary 3D rotation. s_m represents (as introduced above) the scale of the object. The representation of an object from any view is given by a collection of several histograms distributed over all possible views. [9] examines the number of histograms which are needed

for the representation of a 3D object. We concluded from the experiments that, because of the robustness of multidimensional receptive field histograms to view point changes, already a small number of histograms are sufficient in order to obtain high recognition rates.

2.1 Histogram matching for object recognition

The paper [7] experimentally evaluates different histogram matching function for the recognition of objects. This section describes the χ^2–statistics, which has been shown to be the most reliable (in terms of recognition and robustness) in most of the experiments. Histogram matching is not the only possibility to use the developed statistical representation for objects. In [8] we have developed a probabilistic object recognition approach which is based on single measurements in the image and which is entirely based on multidimensional receptive field histograms. Experiments showed that a relatively small portion of the image is sufficient in order to recognize objects of a database of 100 objects.

The χ^2–statistics for the comparison of two histograms is defined by [5]:

$$\chi^2(D(M|I), H(M|o_n, r_i, s_m) = \sum_k \frac{[H(m_k|o_n, r_i, s_m) - D(m_k|I)]^2}{H(m_k|o_n, r_i, s_m) + D(m_k|I)} \qquad (4)$$

where $D(M|I)$ is the histogram of the test-image I of an (unknown) object and $H(M|o_n, r_i, s_m)$ is a histogram of the database. From the comparison of the test histogram $D(M|I)$ of the test-image I to the database of histograms $H(M|o_n, r_i, s_m)$ we obtain an estimate of the object ($\hat{o_n}$) and its pose ($\hat{r_i}$, $\hat{s_m}$) in the test-image:

$$(\hat{o_n}, \hat{r_i}, \hat{s_m}) : \min_{n,i,m} \chi^2(D(M|I), H(M|o_n, r_i, s_m)) \qquad (5)$$

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. This analogy is most appropriate for a statistical object representation as introduced in the previous section. But the analogy can be applied to a wide variety of object recognition processes. The only assumption for the validity of the analogy is that the

"messages" (or measurements) m_k obtained from the object are member of a certain finite measurement set M.

The following section 3.1 summarizes basic concepts from information theory (see for example [6]) and gives a first interpretation of the concepts in the context of object recognition. The reader familiar with information theory should move directly to section 3.2 which discusses in more detail the interpretation of different entropies in the context of object recognition. Section 3.3 introduces the transinformation (or mutual information) of the object recognition process. In particular one can apply the transinformation for the evaluation of the employed measurement set. Experiments show that we can predict the performance of the object recognition process using a certain measurement set M. Section 3.4 finally defines the information theoretical concepts of capacity, redundancy and efficiency in the context of object recognition.

3.1 Information measurement

We partition the sample space Ω_X in a finite number of mutually exclusive events x_n, whose probabilities $p(x_n)$ are assumed to be known. The events x_n are a complete partition in the sense that:

$$\bigcup_{n=1}^{N} x_n = \Omega_X \tag{6}$$

$$\sum_{n=1}^{N} p(x_n) = 1 \tag{7}$$

A probability scheme with these properties is called a *complete finite probability scheme*. 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 following well known formula:

$$H(X) = -\sum_{n=1}^{N} p(x_n) \log(p(x_n)) \tag{8}$$

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:

$$H(X) = \overline{I(x_n)} = \sum_{n=1}^{N} p(x_n) I(x_n) \tag{9}$$

The object set $\Omega_O = \bigcup_{n=1}^{N} o_n$ with there probabilities $p(o_n)$ form a complete finite probability scheme. Therefore we can follow formula 8 in order to define the average self–information $H(O)$ of each object o_n:

$$H(O) = -\sum_{n=1}^{N} p(o_n) \log(p(o_n)) \tag{10}$$

The same analogy can be applied to the measurement set M which also forms a complete finite probability scheme:

$$H(M) = -\sum_{k=1}^{K} p(m_k) \log(p(m_k)) \tag{11}$$

In the context of the transmission of information through a channel one needs to know the relation between the input symbols and the output symbols. The following therefore develops information measurement for the two–dimensional case.

Information measurement for the two–dimensional case: The measurement $H(X)$ of *uncertainty* or *information* can be generalized for a two-dimensional discrete finite probability scheme. Such a probability scheme is given by two sample spaces Ω_X and Ω_Y where we select complete event sets x_n and y_k in the sense of equation 6 and 7. 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 the following probability matrix:

$$P(X \wedge Y) = \begin{pmatrix} p(x_1 \wedge y_1) & p(x_1 \wedge y_2) & \dots & p(x_1 \wedge y_K) \\ p(x_2 \wedge y_1) & p(x_2 \wedge y_2) & \dots & p(x_2 \wedge y_K) \\ \vdots & & \ddots & \vdots \\ p(x_N \wedge y_1) & p(x_N \wedge y_2) & \dots & p(x_N \wedge y_K) \end{pmatrix} \tag{12}$$

Therefore we have three complete probability schemes, namely $P(X)$, $P(Y)$ and $P(X \wedge Y)$. We can define three corresponding entropies:

$$H(X) = -\sum_{n=1}^{N} p(x_n)\log(p(x_n)) \tag{13}$$

$$H(Y) = -\sum_{k=1}^{K} p(y_k)\log(p(y_k)) \tag{14}$$

$$H(X \wedge Y) = -\sum_{n=1}^{N}\sum_{k=1}^{K} p(x_n \wedge y_k)\log(p(x_n \wedge y_k)) \tag{15}$$

with

$$p(x_n) = \sum_{k=1}^{K} p(x_n \wedge y_k) \tag{16}$$

$$p(y_k) = \sum_{n=1}^{N} p(x_n \wedge y_k) \tag{17}$$

$H(X)$ is called the marginal entropy of X, $H(Y)$ the marginal entropy of Y and $H(X \wedge Y)$ the joint entropy. One can define two further entropies: the conditional entropies $H(X|Y)$ and $H(Y|X)$. In order to obtain the formula for $H(X|Y)$ we start with the self-information of the event $(x_n|y_k)$ which is:

$$I(x_n|y_k) = -\log(p(x_n|y_k)) \tag{18}$$

The average Information $H(X|y_k)$ given a certain y_k is then:

$$H(X|y_k) = \overline{I(x_n|y_k)} = -\sum_{n=1}^{N} p(x_n|y_k)\log(p(x_n|y_k)) \tag{19}$$

The average conditional entropy $H(X|Y)$ is therefore calculated by:

$$H(X|Y) = \overline{H(X|y_k)} = \sum_{k=1}^{K} p(y_k)H(X|y_k) \tag{20}$$

$$= -\sum_{k=1}^{K} p(y_k)\sum_{n=1}^{N} p(x_n|y_k)\log(p(x_n|y_k)) \tag{21}$$

$$= -\sum_{k=1}^{K}\sum_{n=1}^{N} p(x_n \wedge y_k)\log(p(x_n|y_k)) \tag{22}$$

Analogous we can derive the formula for the conditional entropy $H(Y|X)$:

$$H(Y|X) = -\sum_{k=1}^{K}\sum_{n=1}^{N} p(x_n \wedge y_k)\log(p(y_y|x_n)) \tag{23}$$

3.2 Application of information theory to object recognition

In the previous section we derived five entropies for a two-dimensional discrete finite probability scheme without giving their interpretation. In information theory a two-dimensional 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 matrix 12 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 one extracts 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 matrix (cf. formula 12). Using this analogy between a communication network and the object recognition process one can interpret the five entropies $H(O)$, $H(M)$, $H(O \wedge M)$, $H(O|M)$ and $H(M|O)$ in the following way:

- $H(O)$ (formula 13) is the average information of each object o_n,
- $H(M)$ (formula 14) is the average information of each measurement m_k,
- $H(O \wedge M)$ (formula 15) is the overall uncertainty of the recognition process,
- $H(O|M)$ (formula 22) gives an indication of the average "noise" or error of the recognition process,
- $H(M|O)$ (cf. formula 23) indicates the overall quality of the recognition process. The smaller $H(M|O)$ the better the object set O can be recognized with the measurement set M.

Since we consider that all objects o_n are equally probable ($p(o_n) = 1/N$) the entropy $H(O)$ is constant: $H(O) = \log(N)$.

The remaining four entropies will change significantly, when we change the measurement set M. Perhaps the most interesting entropy is $H(O|M)$ which indicates the error of the recognition process. In order to show the influence of the chosen measurement set M we will develop the formula of $H(O|M)$ (cf. formula 22):

$$H(O|M) = -\sum_{k=1}^{K}\sum_{n=1}^{N} p(o_n \wedge m_k)\log(p(o_n|m_k))$$

$$
\begin{aligned}
&= -\sum_{k=1}^{K}\sum_{n=1}^{N} p(o_n \wedge m_k) \log \frac{p(o_n \wedge m_k)}{p(m_k)} \\
&= -\sum_{k=1}^{K}\sum_{n=1}^{N} p(o_n \wedge m_k) \log(p(o_n \wedge m_k)) + \sum_{k=1}^{K} p(m_k) \log(p(m_k)) \\
&= H(O \wedge M) - H(M) \qquad (24)
\end{aligned}
$$

When one minimizes $H(O|M)$, one reduces the average error of the recognition process. Following formula 24 this can be obtained by minimizing the overall uncertainty $H(O \wedge M)$ and by maximizing $H(M)$. $H(M)$ has its maximum when all measurements m_k are equally-probable. $H(M \wedge O)$ is minimal, when for each object o_n there exist exactly one measurement m_k so that $p(o_n|m_k) = 1$ and $p(m_k|o_n) = 1$ (this corresponds to the case that each row of matrix 12 contains at exactly one position the value $1/N$). More general one can say the smaller $H(M \wedge O)$ the more significant are the measurements m_k in average. Therefore formula 24 provides a possibility to compare numerically different measurement sets m_k for the same object set o_n. Formula 24 can be applied in the context of many object recognition processes, as soon one can calculate or estimate the joint probability matrix 12.

3.3 Transinformation of the object recognition process

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 one can calculate the transinformation of an object/measurement pair (o_n, m_k):

$$
T(o_n, m_k) = \log \frac{p(o_n \wedge m_k)}{p(o_n)p(m_k)} \qquad (25)
$$

Therefore the average transinformation per symbol pair is:

$$
T(O, M) = \overline{T(o_n, m_k)} = \sum_{n=1}^{N}\sum_{k=1}^{K} p(o_n \wedge m_k) \log \frac{p(o_n \wedge m_k)}{p(o_n)p(m_k)} \qquad (26)
$$

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. By applying the definition (formula 26) one can easily show that:

$$T(O,M) = H(O) + H(M) - H(O \wedge M) \quad (27)$$
$$= H(O) - H(O|M) \quad (28)$$
$$= H(M) - H(M|O) \quad (29)$$

Therefore if one wants to maximize the information transmitted by the channel (respectively by the recognition process) one should minimize $H(O|M)$ (we assume $H(O)$ to be constant as mentioned above) (cf. formula 28). Or by using formula 29 one can say, that the maximization of $H(M)$ and the minimization of $H(M|O)$ result in the maximization of the transinformation. Similar results have been obtained from the analysis of formula 24.

Using the fact that $H(O \wedge M) \geq \max(H(O), H(M))$ in formula 27 one obtains the following bound of the transinformation $T(O,M)$:

$$T(O,M) \leq \min(H(O), H(M)) \quad (30)$$

Since we consider $H(O)$ to be constant (all objects o_n are equally probable), the upper bound of the transinformation $T(O,M)$ can be increased by $H(M)$ until $H(M)$ equals $H(O)$.

The most interesting idea of using transinformation is the possibility to compare different measurement sets M for a certain object set O. By applying the transinformation only onto a subset of O one can obtain the most appropriate measurement set M for the discrimination of objects of that subset.

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. Figure 1 shows the result of this evaluation. In this graph three different measurement combination are used: *Dx-Dy* corresponds the first Gaussian derivative in $x-$ and in $y-$direction. *Mag-Lap* is corresponds also to a two–dimensional measurement vector, namely the magnitude of the first Gaussian derivative and the Laplace operator. *Dx-Dy-Lap* finally corresponds to a three-dimensional measurement vector of the first derivatives in $x-$ and $y-$direction and the Laplace operator (we use the usual definition of Gaussian derivatives which we describe in [8]).

In figure 1 one can see that the choice of the measurement set plays an important role for the transinformation of the recognition process. The graph shows a significant increase of the transinformation for the three-dimensional measurement set *Dx-Dy-Lap* relative to the two-dimensional 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 expected to be even

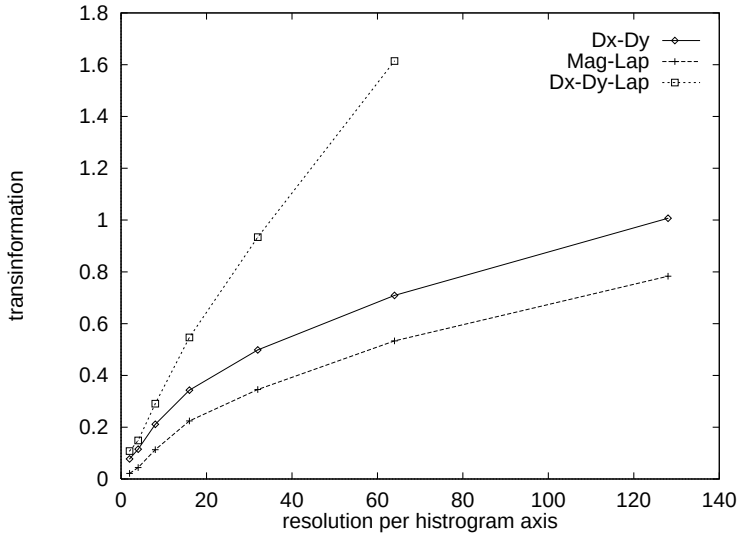

Fig. 1. The transinformation for 100 objects and for different measurement sets: *Dx-Dy* and *Mag-Lap* correspond to two-dimensional histograms and *Dx-Dy-Lap* corresponds to a three dimensional histogram. The horizontal axis shows different resolutions for the histogram axis and the vertical axis show the transinformation for a particular resolution

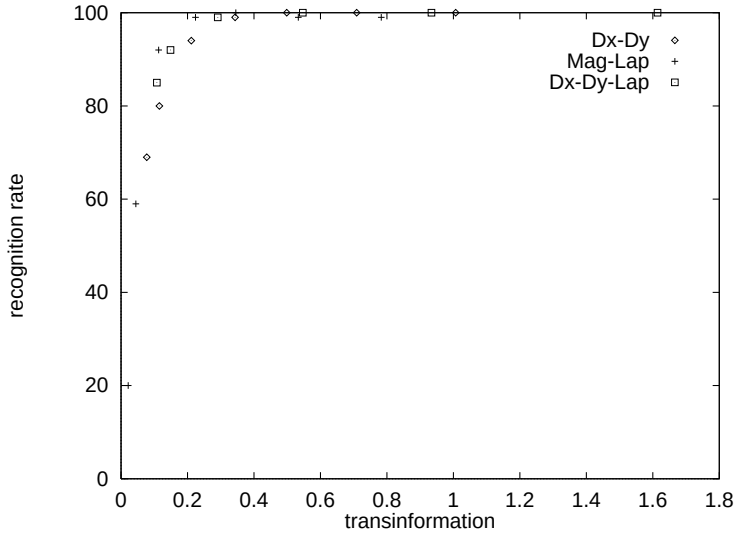

Fig. 2. The relation between the transinformation for 100 objects (for different measurement sets, cf. figure 1) and the recognition rate of 100 test images. One observes that all points lie on approximately the same curve, which shows the strong relation between the transinformation and the recognition rate. Therefore the transinformation can be used for the prediction of the recognition rate

more important by adding more independent dimensions. It is interesting to note that the result of *Dx-Dy* and *Mag-Lap* are qualitatively similar. Nevertheless *Dx-Dy* gives better results than *Mag-Lap*. This can be explained by the fact that *Dx-Dy* preserves the orientation information of the first derivative whereas *Mag-Lap* is rotationally invariant.

Figure 2 shows the relation between the transinformation (cf. figure 1) and the recognition rate for the different measurement sets. We used the same measurement sets and the same database images as in figure 1. As test images we used 100 images with a view point change of 5° relative to the database images. The figure shows the strong relation between the transinformation and the recognition rate. We are therefore convinced that one can predict the quality of recognition based on the transinformation of the object recognition process.

3.4 Capacity, redundancy and efficiency of the object recognition process

For completeness we will apply the information theoretical concepts of capacity, redundancy and efficiency to the introduced analogy to the object recognition process.

In information theory one defines the channel capacity as the maximum of the transinformation:

$$Capacity = \max_{P(O)} T(O,M) = \max_{P(O)} \left(H(M) - H(M|O)\right) \tag{31}$$

The maximization is with respect to all possible sets of probabilities $P(O)$ of the input, that means of the objects o_n. Therefore the capacity is calculated for a specific probability scheme $p(m_k|o_n)$.

Absolute redundancy is defined as the difference of the $Capacity$ and the transinformation:

$$Redundancy_{absolut} = Capacity - T(O,M) \tag{32}$$

Relative redundancy is defined as the relation of $Redundancy_{absolut}$ to $Capacity$:

$$Redundancy_{relative} = \frac{Capacity - T(O,M)}{Capacity} \tag{33}$$

Efficiency finally is defined as $Effiency = 1 - Redundancy_{relative}$.

4 Choice of the most significant viewpoints

In the previous section we showed the application of information theoretical concepts to the process of object recognition. In this section we want to go a step further in applying the developed analogy in the context of viewpoint planning for object recognition. More precisely we will calculate the transinformation of each viewpoint of an object in order to choose the most "significant" viewpoints of an object. Section 5 will use these viewpoints in an active object recognition approach.

4.1 Transinformation of a single viewpoint of an object

In section 3.3 we defined the transinformation of the object recognition process as a whole based on the event pairs (o_n, m_k). In the following we will develop a formula for the choice of the most significant viewpoint of an object. Starting from the original formula 26 we can rewrite the transinformation:

$$T(O,M) = \sum_{n=1}^{N}\sum_{k=1}^{K} p(o_n \wedge m_k) \log \frac{p(o_n \wedge m_k)}{p(o_n)p(m_k)} \tag{34}$$

$$= \sum_{n=1}^{N} p(o_n) \sum_{k=1}^{K} p(m_k|o_n) \log \frac{p(m_k|o_n)p(o_n)}{p(o_n)p(m_k)} \tag{35}$$

$$= \sum_{n=1}^{N} p(o_n) \sum_{k=1}^{K} p(m_k|o_n) \log \frac{p(m_k|o_n)}{\sum_n p(m_k|o_n)p(o_n)} \tag{36}$$

The transinformation can be interpreted as the average transinformation for an object o_n which we want to define as:

$$T(o_n, M) = \sum_{k=1}^{K} p(m_k|o_n) \log \frac{p(m_k|o_n)}{\sum_i p(m_k|o_i)p(o_i)} \tag{37}$$

The probability $p(m_k|o_n)$ corresponds to the "complete" probability density function as introduced before. Since we want to calculate the transinformation of each individual viewpoint of an object we want to define the transinformation of an object (o_n) at a certain pose $(r_i,\ s_m)$ as:

$$T(o_n, r_i, s_m, M) = \tag{38}$$
$$\sum_{k=1}^{K} p(m_k|o_n, r_i, s_m) \log \frac{p(m_k|o_n, r_i, s_m)}{\sum_{n,i,m} p(m_k|o_n, r_i, s_m)p(o_n, r_i, s_m)}$$

5 Active object recognition

This section describes an active object recognition algorithm. The principal idea of the algorithm is to hypothesize the object identity and its pose from a test-image (single view recognition). Based on this estimate the algorithm moves the camera to the most discriminant viewpoint(s) of the hypothesized object. The information gathered from the new viewing direction is then used for the verification of the hypothesized object as well as its hypothesized pose.

Experimental results (section 5.2) show that the algorithm can resolve all (except one) of the ambiguities which are inherent to a single–view recognition algorithm. Most interestingly the algorithm incorporates 3D information from the statistical representation of the object. Remaining errors (for one verification step) are therefore 3D consistent. By applying multiple verification steps the algorithm incorporates more 3D information which results in a nearly 100% recognition rate.

5.1 Outline of the active object recognition algorithm

The outline of the active recognition algorithm contains three steps:

Hypothesis generation: the hypothesis is generated by matching of the histogram $D(M|I)$ of the test images I with the database of histograms. An object and its pose is hypothesized, if the corresponding multidimensional receptive field histogram produces the minimum of the χ^2-statistics: the hypothesized object is $\hat{o_n}$ with its pose parameters $\hat{r_i}$, $\hat{s_m}$ (cf. formula 5).

Camera movement: the camera is moved to the most (or second most) discriminant viewpoint of the object $\hat{o_n}$. This discriminant viewpoint is calculated off-line on the basis of formula 38. The movement of the camera is calculated on the basis of the difference Δr and Δs between the current estimated pose $\hat{r_i}$, $\hat{s_m}$ and the pose of the most discriminant viewpoint.

Verification of the hypothesis: at the new camera position we will get (cf. formula 5) again a hypothesized object and its hypothesized pose. Different conditions can be employed for the verification step. First of all the hypothesized object should be the same before (time $(t-1)$) and after (time (t)) the camera movement:

$$\hat{o_n}(t) = \hat{o_n}(t-1) \tag{39}$$

Furthermore we can use the knowledge of the camera movement which has been calculated on the basis of the differences Δr and Δs. By using these differences one can predict the observation at the new camera position and compare them to the obtained hypothesis:

$$\hat{r_i}(t) = \hat{r_i}(t-1) + \Delta r(t-1) \pm \epsilon(r) \quad (40)$$
$$\hat{s_m}(t) = \hat{s_m}(t-1) + \Delta s(t-1) \pm \epsilon(s) \quad (41)$$

$\epsilon(r)$, $\epsilon(s)$ correspond to the allowed error in the estimation.

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

The most attractive property of the proposed active object recognition process is that not only 2D information (a single image) but also 3D information of the probability density function of the objects is used in order to verify the hypothesized object. That implies that errors of the proposed approach should be only possible between objects which are 3D–compatible (for example different cubes). in the experiments described below we observe this property for the errors of the system.

If one is interested in an object recognition process with minimal false positive and minimal false negative at the same time, one can introduce several verification steps. If one uses L verification steps one obtains the following conditions for the verification step:

$$\forall t = 1, ..., L : \hat{o_n}(t) = \hat{o_n}(t-1) \quad (42)$$
$$\hat{r_i}(t) = \hat{r_i}(t-1) + \Delta r(t-1) \pm \epsilon(r) \quad (43)$$
$$\hat{s_m}(t) = \hat{s_m}(t-1) + \Delta s(t-1) \pm \epsilon(s) \quad (44)$$

5.2 Experimental results

This section describes an experiment in order to show the applicability of the proposed active object recognition approach. The experiments are based on the Columbia image database of 100 objects [4], which contains 7200 image, 72 for each object (originally the database contains color images which we converted to grey scale images). The images are taken under controlled lightening conditions in front of a black background. Figure 3 shows some of the 100 objects. The 72 views of an object are taken from a fixed camera position whereas the object has been turned in 5° intervals on a turntable.

Unfortunately the database contains only one rotational freedom (which we

Fig. 3. Columbia image database: 20 of the 100 objects

will call α in the following) for the objects. That means that three parameter of the pose estimation (namely β, γ and s) are not considered in the experiment. The reason of using this database was that we can simulate the camera movement by "turning" the object in front of the camera. Therefore we can validate the proposed algorithm without the dependency onto the camera movement and the camera calibration (in order to move the camera relative to the object one has to calibrate the camera at least approximately).

Half of the images (every 10°) of the objects ($100 \times 36 = 3600$ images) are used as database. The remaining half of the images are taken as test set. For each of the database images we calculate its histogram of the first Gaussian derivative in x– and in y–direction. In this particular experiment we used a resolution of 16 cells per histogram axis so that each of the 3600 images is represented by $16^2 = 256$ numbers. For each of the objects we have calculated the most and the second most discriminant viewpoint.

Table 1 shows the results which we obtained by the application of the active object recognition algorithm introduced above. The first row shows the number of verification steps which we used in order to accept a hypothesized object. Without using any verification (first line of the table) we obtain 59 misclassification of the 3600 test images. By using only one verification step we can reduce this number to 31 misclassification. By using 2 verification steps we can resolve nearly all ambiguities and we obtain only one misclassification. These results validate the applicability of the proposed active recognition approach. (During the verification steps we allowed an error $\epsilon(r)$ of up to 10°, cf. formula 43).

verification steps	% recognition	number of errors
0	98.36	59
1	99.14	31
2	99.97	1

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

Fig. 4. The 5 objects which are confused by the recognition algorithm with one verifications step

It is interesting to note that all misclassification produced with one verification step are 3D consistent. Figure 4 shows the five objects which are not correctly classified but as one of the other four objects. Most errors (20 of 31) 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. Therefore the errors are not arbitrary but rather "systematic". One can discriminate these objects by using color information or other appropriate measurements (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 4.

6 Conclusion

This article motivates the statistical representation of 3D objects by a collection of multidimensional receptive field histograms. Each of the histogram represents a certain appearance of the object. Based on such a statistical representation the article expresses object recognition 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. Following the developed analogy one can also define the capacity, the redundancy and the efficiency of the object recognition process.

Based on the analogy between the transmission of information through a channel and an object recognition process one can calculate the *transinformation* of the recognition process. Experiments show the applicability and the validity of

the transinformation for the quantitative evaluation of measurement (or feature) sets for object recognition. 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. Even more interestingly the algorithm incorporates 3D information of the objects entirely based on 2D measurements in images of the object.

References

[1] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 13(9):891–906, 1991.

[2] J. Hornegger and H. Niemann. Statistical learning, localization and identification of objects. In *ICCV'95 Fifth International Conference on Computer Vision*, pages 914–919, 1995.

[3] J. L. Mundy and Andrew 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, Departement of Computer Science, Columbia University, 1996.

[5] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, 2nd edition, 1992.

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

[7] B. Schiele and J. L. Crowley. Object recognition using multidimensional receptive field histograms. In *ECCV'96, Fourth European Conference on Computer Vision, Volume I*, pages 610–619, 14–16 April 1996.

[8] B. Schiele and J. L. Crowley. Probabilistic object recognition using multidimensional receptive field histograms. In *ICPR'96, Proceedings of the 13th International Conference on Pattern Recognition, Volume B*, pages 50–54, August 1996.

[9] B. Schiele and J. L. Crowley. The robustness of object recognition to view point changes using multidimensional receptive field histograms. Presented at *ECIS-VAP meeting, Object Recognition Day*. Available via WWW [2], March 1996.

[2] http://pandora.imag.fr/Prima/schiele/