M.I.T. Media Laboratory Perceptual Computing Section Technical Report No. 499
to appear in ICCV'99

Probabilistic Object Recognition and Localization

Bernt Schiele and Alex Pentland

The Media Laboratory, Room E15-384c,
Massachusetts Institute of Technology
20 Ames St., Cambridge, MA 02139
email: {bernt,sandy}@media.mit.edu

Abstract

The appearance of objects consists of regions of local structure as well as dependencies between these regions. The local structure can be characterized by a vector of local features measured by local operators such as Gaussian derivatives or Gabor filters. This paper presents a technique in which the appearance of objects is represented by the joint statistics of local neighborhood operators. A probabilistic technique based on joint statistics is developed for the identification of multiple objects at arbitrary positions and orientations. Furthermore, by incorporating structural dependencies, a procedure for probabilistic localization of objects is obtained. The current recognition system runs at approximately 10Hz on a Silicon O2. Experimental results are provided and an application using a head mounted camera is described.

1 Introduction

Recently, various successful approaches for object recognition have been proposed using object representation schemes based on local image descriptors. The underlying idea of these approaches is that vectors of local image descriptors, such as Gaussian derivatives, contain highly discriminating information and therefore can be used for the representation as well as the recognition of objects. The main advantage of these techniques is that local image descriptors can be extracted reliably from images. The recognition algorithms can be therefore robust to appearance changes of the objects caused by different imaging conditions or viewpoints.

Approaches based on local image measurements differ mainly on how structural dependencies between local measurements are coded and how they are used for recognition. Rao and Ballard [11] e.g. use a high–dimensional "iconic" feature vector. The feature vector is defined over a fixed global grid, consisting of 45 dimensional vectors of Gaussian derivatives at each grid point. Recognition is attained by matching the iconic feature vector of the image with the database vectors. Structural dependencies are coded by means of the global grid. Schmid and Mohr [13] use a vector of 2D rotation invariant gray level derivatives as indices into a hash-table. The number of entries into the hash-table are reduced by the use of an interest point detector. A standard voting scheme is used for the recognition of objects. In order to reduce the number of false positives (a standard problem of hash-table based approaches [4]) affine invariants of point-couples are employed. Structural dependencies are therefore coded only locally and are inherent in the hash-table.

Other approaches, i.e. Swain and Ballard [15], Mel [7] and Schiele and Crowley [12] propose the representation of objects by the global statistics of their local image measurements. These techniques determine the most probable object in the scene by matching the statistics from a region of the image with the statistics from a sample of the object. Here, structural dependencies are not coded explicitly. The matching is obtained by means of simple histogram intersection or by more statistically motivated measures [1] such as the χ^2–statistics. The main differences are the employed image measurements: Swain and Ballard use a three-dimensional color vector; Mel employs 102 feature "channels" that are sensitive to contour, texture and color cues; and Schiele and Crowley use multi-dimensional vectors of Gaussian derivatives.

The above mentioned approaches take advantage of the fact that properly chosen local image measurements contain enough information for the discrimination of a large number of objects. However, many approaches only use global representations of structural dependencies, making it difficult to apply them in the presence of significant partial occlusions or object deformations. Hash-table based approaches, on the other hand, are strictly local since they only code local dependencies between image measurements. Due to the locality of the representation, the global structure cannot be used for recognition.

This paper proposes a hybrid representation of ob-

jects which profits from both the discriminant power of local image measurements and the coding of structural dependencies. The representation is flexible in the sense that it can be adapted online during recognition depending on the current context (as a priori knowledge or context information).

The reminder of the paper is organized as follows: the next section introduces the representation of objects by means of global statistics over a set of local image measurements. In section 3 a probabilistic algorithm for the recognition of objects in the presence of significant partial occlusion is developed and tested. Section 4 extends this approach to a flexible representation of objects including the local statistics of image measurements as well as structural dependencies between them. A probabilistic procedure for the localization is developed and the robustness and the accuracy of the procedure are analyzed. A real-time application of the system using a head-mounted camera is described in section 5.

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 [owncitation] 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. Possible 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:

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

Dimensionality reduction

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 [8] 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 [owncitation].

It is difficult to model *partial occlusion* in a general way. Sections 3 and 4 propose a probabilistic object recognition approach which recognizes objects by the observation of only a small portion of the object. This makes 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 (see discussion below). 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.

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 [13] (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 (equation 1) are three (or two) components of the rotation R and one component of the translation (rep-

resented by the size σ of the object), called the pose $s = (\sigma, R)$:

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

Representation of $p(M|o_n, s)$

The probability density can be represented in various ways such as by means of a mixture of multivariate Gaussians [5] or kernel estimator [14, 10]. In this paper we use multidimensional histograms. The following shortly discusses why these histograms can be estimated reliably.

The six-dimensional histograms used in the following contain about 5000 different cells which have to be estimated. In order to obtain a reliable estimate the number of samples should be at least the same magnitude as the number of cells or more [3, 14] (even though it is difficult to access in general the amount of data necessary). Since every image pixel corresponds to one sample and we are using images of 400x400 pixels the number of samples is sufficiently large for a reliable estimate. By additionally using a Laplace-prior (uniform prior) low-density regions are "filled" such that the overall estimate becomes biased but is still reliable. We have experimented with different weighting between the prior and the data and observed that the results are relatively insensitive to the exact amount of bias (see [owncitation] for further details). The use of kernel estimators (i.e. approximated by smoothed histograms) is a way to increase the ability to generalize from data. Using a kernel estimator, however, did not alter the overall results significantly. Therefore, the experimental results in the following are obtained using multidimensional histograms providing good discrimination between a large number of objects. However, we envision a more compact representation in the future.

3 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 the pose s_l given a certain local measurement m_k. This probability $p(o_n, s_l|m_k)$ can be calculated by Bayes' rule:

$$p(o_n, s_l|m_k) = \frac{p(m_k|o_n, s_l)p(o_n, s_l)}{p(m_k)} \tag{3}$$

with $p(o_n, s_l)$ the a priori probability of the object o_n at pose s_l, $p(m_k|o_n, s_l)$ the probability density function of object o_n at s_l, and $p(m_k)$ the a priori probability of the image measurement m_k. Having K independent local measurements $m_1, \ldots, m_K$ we can calculate the probability of each object o_n and each pose s_l by:

$$\begin{aligned} p(o_n, s_l|m_1, \ldots, m_K) &= \frac{p(m_1, \ldots, m_K|o_n, s_l)p(o_n, s_l)}{p(m_1, \ldots, m_K)} \\ &= \frac{\prod_k p(m_k|o_n, s_l)p(o_n, s_l)}{\prod_k p(m_k)} \end{aligned}$$

The probabilities $p(m_k|o_n, s_l)$ are directly given by the multidimensional receptive field histograms. Therefore, the above equation shows a calculation of the probability for each object o_n only based on the multidimensional receptive field histograms of the N objects. To guarantee the independence of the different local measurements the minimal distance $d(m_{k_1}, m_{k_2})$ between two measurements m_{k_1} and m_{k_2} must be sufficiently large. In the experiments described below, we choose the minimal distance $d(m_{k_1}, m_{k_2}) \geq 2\sigma$ which is sufficient to guarantee independence of the vectors from a signal processing point of view.

It is worth pointing out several properties of the above equation: the image measurements m_k can be made at arbitrary image positions; it is not necessary to have measurements at special or corresponding points. In other words, the correspondence problem does not have to be solved for recognizing objects. This is a very powerful property since most object recognition algorithms rely on the computationally expensive calculation of correspondence. Secondly, this equation allows the calculation of probabilities with a complexity which is linear in the number of image measurements m_k. More precisely the complexity is $O(NK)$ for the calculation of the probabilities with N the number of objects (enabling the real-time implementation in section 5).

Experimental Results

This section describes an experiment using 103 different objects. Figure 1 shows some of the database objects. More precisely the database contains 20 objects from different viewpoints (Columbia image database [9]) and images of 83 objects comprising image plane rotations and scale changes. In the experiment we use a six–dimensional probability distribution of the filter combination Dx-Dy (first Gaussian derivative in x and y directions) at three different scales with $\sigma_1 = 2$, $\sigma_2 = 2\sigma_1$ and $\sigma_3 = 4\sigma_1$.

We use one image from each of the 83 objects as training images, and calculate pdfs corresponding to different image plane rotations and scale changes using the steerability of Gaussian derivatives to image plane rotation and their adaptability to scale. In order to cover the range of pose changes in the test images, pdfs corresponding to different rotations, namely for the angles ($\alpha = 0°, 20°, \ldots, 340°$) as well as to different scales

Figure 1: 7 of the 103 objects used in the experiments

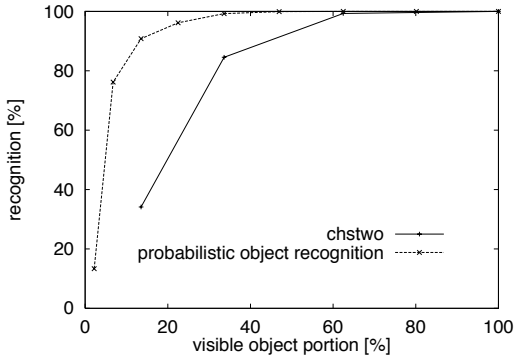

Figure 2: Object recognition in the presence of partial occlusion of 1327 images of 103 different objects

($\sigma = 1.48, 1.7, 2.0, 2.26, 2.62, 3.0$) have been calculated. For the 20 objects of the Columbia image database we used half of the 1440 images as training set. The angle between the different viewpoints is therefore 10^o. The test-set contains 1327 independent test images total, including the remaining half of the Columbia image database (770 images), 18 different image plane rotations for 22 objects (covering 360 degrees rotation), 6 different scales for 30 objects (the overall scale factor is approximately 2.2) and an additional 31 images of the remaining 31 objects including minor pose changes of the objects.

For the experiments we assume that all objects are equi-probable and have probability $p(o_n, s_l) = \frac{1}{NM}$, where N is the number of objects and M the number of poses per object. Furthermore, we only report the precision of the recognition of objects and ignore the estimate of the object's pose.

Figure 2 shows the results obtained by the probabilistic recognition as well as the results obtained by matching the statistics directly by means of the χ^2-statistics. In both cases, 62% of the object needs to be visible for the correct recognition of all 1327 test images. The probabilistic recognition outperforms the matching technique for more important partial occlusions: with 33.6% visibility, the recognition rate is above 99% (10 errors in total). Using 13.5% of the object, the recognition rate is still above 90%. This can be explained by the fact that each single vector contains discriminant information. This point is reinforced by noting that a recognition result of approximately 13% is obtained with only a single measurement vector. The difference in the recognition performance of the two recognition algorithms is explained by the fact that the χ^2-statistics is a global method for matching statistics. The probabilistic recognition algorithm is strictly local since it is based on individual image measurements only.

We can conclude that the proposed probabilistic recognition approach is capable of discriminating 103 objects in the presence of significant scale changes, image plane rotation and viewpoint changes. As mentioned earlier, the recognition results have been obtained without any correspondence between points in the test images and the database. The approach has been shown to be remarkably robust with respect to partial occlusion since a small portion of the object is sufficient in order to obtain good object hypotheses. However, the algorithm does not give an estimate of the position of the object, generally referred to as the localization problem.

4 Probabilistic localization of objects

Figure 3 shows an example of a cluttered scene with multiple objects from our database which are partially visible. The respective visibility of the objects is roughly between 25% and 50%. By applying the algorithm of the previous section locally on different parts of the image we are capable to collect enough evidence in order to identify which objects are present in the scene. In this example all five objects in the image have been identified. However, the algorithm *cannot* estimate the position of the objects (identified by the crosses) in the image since no correspondence is calculated and no position information is used in the statistical representation of the objects[1]. The goal of this section is to estimate the position of the objects and to draw crosses (like the ones shown in the figure) at the corresponding position in the image.

This section proposes a probabilistic technique which incorporates structural dependencies between local regions in a flexible way in order to estimate the object's identity at the same time as the object's position (as indicated by the crosses in figure 3). Probably the most intuitive way to include (local and global) structure into the probability distribution is to model explicitly the position x_j of the object in the image:

[1] A rough localization can be obtained by storing the location of the different image parts with high evidence for the objects

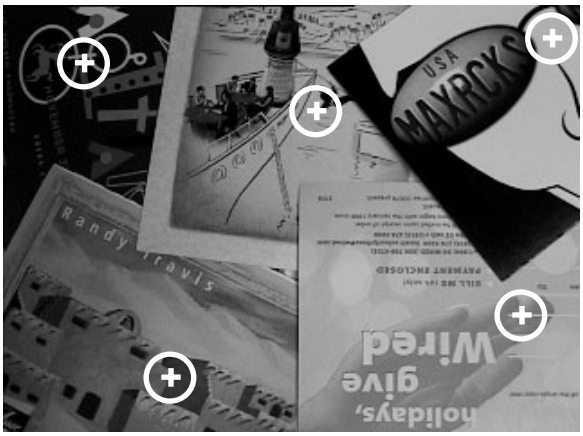

Figure 3: An example of a cluttered scene with multiple objects

$$p(o_n, s_l, x_j | m_k) = \frac{p(m_k | o_n, s_l, x_j) p(o_n, s_l, x_j)}{p(m_k)}$$

Since x_j corresponds to the position of the object $p(m_k | o_n, s_l, x_j)$ can be interpreted as the local pdf of the measurements for a portion of the object. By estimating pdfs for local portions of the object and applying the algorithm of the previous section, we obtain an estimate not only for the object's identity but also for the position of the object in the image. The algorithm of the previous section could be used for the recognition of object portions rather than entire objects. The complexity of this algorithm is $O(NKJ)$ where J is the number of local regions per physical object. This formulation, however, does not use the positional relationships between the different object parts.

In order to incorporate the structural dependencies of local object portions, the relative position of the local pdf's have to be taken into account. Let's consider two image measurements m_{k_1} and m_{k_2} from two arbitrary image locations:

$$p(o_n, s_l, x_j | m_{k_1}, m_{k_2}) = \frac{p(m_{k_1} | m_{k_2}, o_n, s_l, x_j) p(m_{k_2} | o_n, s_l, x_j) p(o_n, s_l, x_j)}{p(m_{k_1} | m_{k_2}) p(m_{k_2})} \quad (4)$$

The structural dependency is represented by the conditional probability $p(m_{k_1} | m_{k_2}, \ldots)$. When the two measurements are sufficiently distant we can assume the independence of the them. Generally, we have to use different local pdfs depending on the relative position of the two measurements. Therefore, by probabilistically choosing local pdf's (depending for example on a priori knowledge of the expected transformations and deformations of the object) the conditional dependency is modeled. The generalization to K image measurements $m_1, m_2, \ldots, m_K$ is straightforward but due to its notational complexity omitted. The complexity of the recognition algorithm is the same as before: $O(NKJ)$.

We'd like to point out the similarity of the proposed algorithm to the voting scheme used by the Generalized Hough Transform or by a hash-table based approach. The typical voting scheme consists of estimating some parameters based on local information and voting for these variables. Here the parameters are the identity and the position of the object. The "votes" are the probabilities of these parameters given the image measurements. The sum of the votes is replaced by the product of the probabilities. Therefore the proposed technique can be seen as a probabilistic voting scheme, which is how the experimental results reported below are obtained.

In order to stress the flexibility of the proposed approach we start with the observation that the recognition algorithm of the previous section is a special case of equation 4: by combining the local pdfs we obtain one global pdf of an object o_n. The conditional probability is thereby dropped and no structural information is used during recognition. The opposite extreme is the case that a local pdf corresponds to one single image measurement m_k. In this case all structural dependencies are preserved but the estimation of the local pdfs generally becomes infeasible (due to lack of data). However, most hash-table based approaches use such an approach by assuming some predefined distribution of the image measurements in the feature space. Schmid and Mohr [13] e.g. use the Mahalanobis distance for matching image measurements and therefore effectively assume a Gaussian distribution of the image measurements.

In most circumstances the most appropriate representation for the recognition of objects is in between these two extreme cases. Typically, it is difficult to predict which granularity of local pdf's will be the most appropriate for recognition in a future unknown context. As already mentioned, local pdf's can be easily combined to construct more global ones enabling a reliable estimate of the object's identity as demonstrated in the previous section.

The benefit of using local pdfs is to obtain an estimate of the object's position as well as an estimate of the object's identity. The finer the granularity of the local pdfs the more precise the position estimation can be. However, there is a tradeoff between the granularity of the pdfs and the robustness of the estimation of the

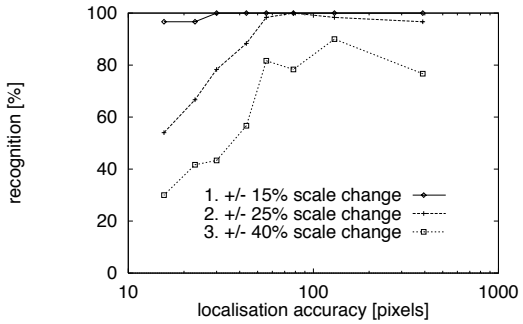

Figure 4: Probabilistic recognition for 30 objects with local pdfs alone as a function of localization accuracy for three test sets with different scale changes.

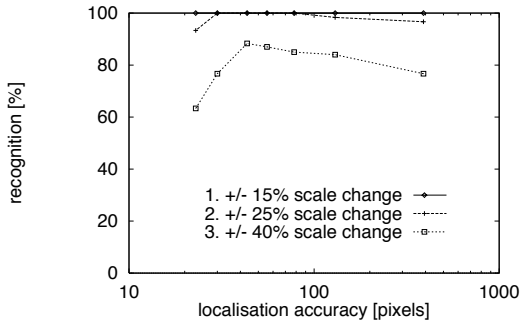

Figure 5: Probabilistic recognition for 30 objects with a mixture of local pdfs as a function of localization accuracy for three test sets with different scale changes.

pdfs which we will analyze in the following. More specifically we propose two different methods to estimate the local pdfs and evaluate their respective robustness.

Experimental Results

We have applied the proposed localization algorithm on the scene shown in figure 3. As for the previous algorithm we could identify all objects in the image. However, the localization procedure also allows to estimate the position of the objects. The crosses in the figure correspond to the estimated positions and the circles around the crosses show the estimated localization accuracy.

The following examines in more detail the robustness and accuracy of the localization procedure. In particular we examine two different methods for the estimation of the local pdfs. The parameter estimated here is the x/y-position of the objects in the image plane. In the following experiment we use a subset of 30 objects from the previous experiment. We calculate pdfs for each object corresponding to one particular scale of the object. In order to test the robustness of the estimation of local pdfs we use three sets of test images compromising $\pm 15\%$, $\pm 25\%$ and $\pm 40\%$ scale changes respectively[2].

A first method of estimating local pdfs is based on the corresponding local object portions only. Figure 4 shows the robustness of the recognition results using this method for the three sets of test images as a function of the granularity of the local pdfs. Note that the granularity of the pdfs is directly related to the localization accuracy. For the first and second test set the recognition accuracy increases as the localization accuracy goes from 400 pixels to about 80 pixels. This is explained by the fact that the use of the additional position information helps to discriminate the objects. Beyond 80 pixels localization accuracy however (more precisely between 15 and 60 pixels) the recognition rate drops drastically for the second test set due to the difficulty to estimate the local pdfs reliably from small object portions. A similar effect can be observed also for the third test set. The curves therefore show the tradeoff between attainable localization accuracy and robustness of estimation for this first method of estimation.

In order to increase the reliability of the estimate of local pdfs the second method uses a hierarchy of pdfs of coarse to fine granularities. The local pdfs are estimated by a mixture of pdfs at different levels of the hierarchy. This method is similar to "shrinkage", a method which has been used in the context of text classification [6]. The mixture coefficients can be learned using a simple form of the EM algorithm. Since we have only one training image per object we use constant mixture coefficients (the coefficients used weight the local pdf the most and decrease exponentially with coarser granularity of the pdfs). Essentially, the reliability of the estimation is increased by biasing the local pdf estimate with neighboring and more global pdfs.

Figure 5 shows the robustness of this second estimation technique for the same granularities of local pdfs as in the previous experiment. This method of pdf mixture generally obtains better results (in particular for the second test set and also for the third) and is overall more stable due to the more reliable pdf estimation. As expected, the gain is larger for the smaller pdfs. A higher precision of the position estimate is therefore obtained since the mixture of pdfs enables the use of local pdfs with a finer granularity. In particular for the second test

[2] These test images could be recognized perfectly by using pdfs which are generalized to different scales covering a range of $\pm 40\%$. In order to test the reliability of the estimation of local pdf we only use pdfs at one particular scale.

set the maximum recognition rate is obtained for a localization accuracy between 30 and 80 pixels. The size of the circles in figure 3 correspond to this localization accuracy.

5 Real-time application using a head-mounted camera

DyPERS, 'Dynamic Personal Enhanced Reality System', [owncitation] is a wearable system that uses augmented reality and computer vision to autonomously retrieve 'media memories' based on associations with real objects that the user encounters. These are evoked as audio and video clips relevant to the user and overlayed in a head mounted display on top of real objects that the user encounters. The user's visual and auditory scene is stored in real-time by the system (upon request) and is then associated (by user input) with a snap shot of a visual object. The object acts as a key such that when the real-time vision system detects its presence in the scene again, DyPERS plays back the appropriate audio-visual sequence.

The system (see figure 6) uses a head-mounted camera to recognize objects in the users' field of view based on the recognition system described in this paper. The same camera is used to record the visual scene of the user. The current system is fully tetherless with wireless connections allowing the user to roam around a significant amount of space (i.e. a few office rooms). The user dons a Sony GlassTron heads-up display with a semi-transparent visor and headphones. Attached to the visor is an ELMO video camera (with wide angle lens) which is aligned as closely as possible with the user's line of sight. Thus the vision system is directed by the user's head motions to interesting objects. In addition, a nearby microphone is incorporated. The A/V data captured by the camera and microphone is continuously broadcast using a wireless radio transmitter. This wireless transmission connects the user and the wearable system to an SGI O2 workstation where the vision and other aspects of the system operate.

The vision system used for this system is a real-time implementation of the probabilistic algorithm described in the previous sections. The recognition system runs at approximately 10Hz on a single Silicon O2 using the OpenGL extension for real-time image convolution. Once an audio-visual clip is stored, the vision system automatically recalls it and plays it back when it detects the object that the user wished to use to remind him of the sequence.

The system was evaluated in a museum-gallery scenario. A small gallery was created in our lab using 20 poster-sized images of various famous works rang-

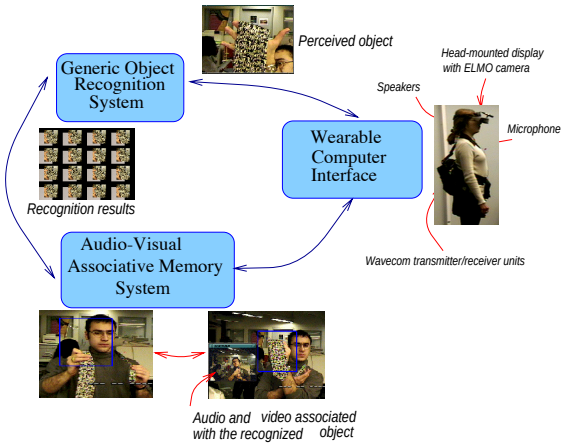

Figure 6: System's architecture of DyPERS

ing from the early 16th century to contemporary art. Subjects were tested in a walk-through of the gallery while a guide was reading a script describing the paintings. After the completion of the tour, the subjects were given a 20-question multiple-choice test containing one query per painting presented. The results indicate that the subjects using DyPERS had an advantage over subjects without any paraphernalia or with standard pencil and paper notes.

Recently, the system has been demonstrated at Nicograph '98 in Tokyo, Japan. Several hundred users tried the system under varying conditions including lighting conditions, occlusion and viewing angles. Users reported that the system recognized all targets perfectly. They were generally impressed with the performance and the short response time of the system.

6 Conclusion

This paper proposes the representation of objects by means of joint statistics over a set of local neighborhood operators. High recognition rates have been reported based only on a small portion of the image, showing the robustness of the approach to significant partial occlusions. Furthermore, the paper proposes a technique to model structural dependencies inherent in the appearance of objects by a set of local joint statistics. This enables not only the recognition of objects but also the estimation of the position of the object in the image. The proposed representation is flexible in the sense that it is possible to choose how much structural dependencies is be used depending on the current context.

A real-time implementation of the system using a head-mounted camera for recognition has been described. This system has been recently used by several hundred users at Nicograph'98.

7 References

[1] M. Basseville. Information: entropies, divergences et moyennes. Technical Report 1020, IRISA, Mai 1996. in French.

[2] M.C. Burl, M. Weber, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *ECCV'98*, pages 628–641, 1998.

[3] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, Inc., 1973.

[4] W.E.L. Grimson and D.P. Huttenlocher. On the sensitivity of the hough transform for object recognition. *PAMI*, 12(3):255–274, March 1990.

[5] J. Hornegger and H. Niemann. Statistical learning, localization and identification of objects. In *ICCV'95*, pages 914–919, 1995.

[6] A. McCallum, R. Rosenfeld, T. Mitchell, and A. Nigam. Improving text classification by shrinkage in a hierarchy of classes. In *International Conference on Machine Learnung*, 1998.

[7] B.W. Mel. Seemore: Combing color, shape, and texture histogramming in a neurally-inspired approach to visual object recognition. *Neural Computation*, 9:777–804, 1997.

[8] J. L. Mundy and Andrew Zisserman, editors. *Geometric Invariance in Computer Vision*. MIT Press, 1992.

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

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

[11] R.P.N. Rao and D.H. Ballard. An active vision architecture based on iconic representations. *Artificial Intelligence*, 78:461–505, 1995.

[12] B. Schiele and J.L. Crowley. Object recognition using multidimensional receptive field histograms. In *ECCV'96, Vol I*, pages 610–619, 1996.

[13] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *PAMI*, 19(5):530–535, 1997.

[14] J.S. Simonoff. *Smoothing Methods in Statistics*. Springer, 1996.

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