Class-based matching of object parts

Evgeniy Bart Shimon Ullman

Department of Computer Science and Applied Mathematics
Weizmann Institute of Science
Rehovot, ISRAEL 76100
{evgeniy.bart, shimon.ullman}@weizmann.ac.il

Abstract

We develop a novel technique for class-based matching of object parts across large changes in viewing conditions. Given a set of images of objects from a given class under different viewing conditions, the algorithm identifies corresponding regions depicting the same object part in different images. The technique is based on using the equivalence of corresponding features in different viewing conditions. This equivalence-based matching scheme is not restricted to planar components or affine transformations. As a result, it identifies corresponding parts more accurately and under more general conditions than previous methods. The scheme is general and works for a variety of natural object classes. We demonstrate that using the proposed methods, a dense set of accurate correspondences can be obtained. Experimental comparisons to several known techniques are presented. An application to the problem of invariant object recognition is shown, and additional applications to wide-baseline stereo are discussed.

1. Introduction

In this paper, we consider the problem of matching corresponding parts of objects across large changes in viewing conditions. The input to the algorithm is a set of images of objects from a given class under different viewing conditions. From this set, the algorithm automatically extracts object parts, and matches corresponding parts in different images. These parts can then be used for automatic image interpretation in terms of objects and their parts, invariant recognition and wide-baseline matching. An example task is to obtain a gallery of face parts such as the eyes, nose, or mouth from face images taken under markedly different viewing directions and illumination, as illustrated in Figure 1.

The problem of generic feature matching has been considered in the past [1–4]. Existing schemes are not restricted to a particular object class. However, they make strong assumptions about the input data. For example, in [1, 5] the availability of video sequences is assumed. Schemes described in [2–4] can handle still images, but assume local scene planarity. When these assumptions are violated, the performance reduces to unacceptable level (see section 4). As a result, the need exists for algorithms that are capable to overcome these restrictions and match reliably non-planar parts under large variations in viewing conditions.

Figure 1: Illustration of the matching problem. Sample input to the system includes images of the same face in several viewing conditions (top row). Desired output includes the images of face parts such as eye (middle row), mouth (bottom row), etc., in all these conditions.

In the scheme described below, we use class-based information in order to obtain reliable part matching. The idea is that in addition to the two images to be matched, examples of objects of the same general class in similar viewing conditions will be used. We will show that this allows to learn the effect of the transformation on the parts of objects of a given class and consequently identify the images of the same object part under much broader conditions than previous schemes.

The motivation for studying class-based part matching arises for several reasons. First, in several popular schemes of object classification [6, 7], indexing and retrieval [3], objects are represented by their constituent parts. The ability to identify the same object part under different viewing conditions would enable these schemes to perform view-invariant recognition [8], which is an important task in computer vision. Second, the ability to reliably identify and localize object parts, in addition to the recognition of the entire object, is of interest on its own right. Third, the problem of feature matching is central to such tasks as tracking and wide-baseline stereo. As discussed in section 6, the use of class-based information for highly familiar objects, such as faces, may improve current tracking and stereo techniques.

The remainder of this paper is organized as follows. In the next section, we review some of the relevant previous approaches to the problem of matching local object parts. In section 3, we describe the proposed class-based matching scheme. In section 4 the performance of the new scheme is compared with several popular matching algorithms. In section 5, an application to the problem of view-invariant object recognition is shown. We conclude with some general remarks in section 6.

2. Previous approaches to matching

Matching features across views requires predicting how the change of viewing conditions will affect the feature's appearance. In simple cases (e.g. when images are related by pure translation) one may assume that the feature's appearance remains constant and only its position changes. This assumption, called 'brightness constancy', is used in several well-known feature tracking and optical flow algorithms [1, 9]. Under the brightness constancy assumption, features can be matched using the minimal sum of squared differences (min-SSD) criterion.

However, in most practical cases the variations of viewing conditions affect the feature's appearance considerably. The impact of viewing conditions is often more significant than the impact of the feature's identity [10].

One general approach to cope with this difficulty is to approximate the transformation of the feature's appearance by some parametric model. Typically, affine or low-order polynomial models of illumination and geometry are used (e.g. [11]), but more complicated schemes [5, 12] have also been proposed. If estimates of the parameters are known, the features may be matched across transformation (e.g. by warping). However, estimating the parameters is often a difficult task. Most schemes [5,12] handle it by using video sequences and updating the parameters for every frame. The task of incremental updating is easier, because the difference in parameters between successive frames is small. The drawback is the necessity to use video sequences, which are not always available, and require additional effort to capture, store and manipulate. In addition, the approximation provided by common parametric models [11] is only valid for a limited range of transformations.

An approach that does not require an estimation of the transformation parameters is to use invariant features. Many popular approaches [2–4, 13] work with so-called affine invariants. The general idea is that the features extracted from an image are normalized with respect to affine transformations. Therefore, features differing by an affine transformation have identical representation and can be matched directly, without the need to estimate the parameters of the transformation. A significant drawback of this approach is that affine approximation holds only for locally planar scenes and affine illumination changes. In practice, many natural objects (such as faces) are not planar. In addition, illumination changes perturb image intensity in a highly non-linear way due to factors such as specularities (highlights) and cast shadows. Invariants more general than affine exist [14], but are usually sparse and therefore insufficient for most tasks. An additional drawback of methods utilizing invariants is the lack of control over the features. Since not all image points are invariant, it is impossible to match a particular point of interest; only the points chosen by the algorithm as invariant can be used.

A scheme that can cope with complex intensity transformations was described in [15]. However, this technique can only handle affine geometric distortions and requires a 3D model for more complex transformations.

In [8], matching pairs of features were learned from video sequences. A similar idea was described in [16], where the average shape of each feature across the transformation was used. However, these methods require examples of correct matches to be provided (in contrast to the scheme proposed here).

3. Class-based matching by fragment equivalence

In this section, we describe the proposed algorithm for class-based matching. In section 3.1, the idea of utilizing equivalence criterion for matching individual object parts is presented. In section 3.2, we describe how the accuracy of matching is improved by exploiting geometric constraints. The final algorithm that combines the appearance and geometry is described in section 3.3.

3.1. Matching object parts

Before describing the part matching method, we first briefly describe relevant aspects of a popular fragment-based object representation scheme [6, 7] that is used by the current method. In these schemes, objects from a general class (such as cars or faces) are represented by their constituent

parts. For example, parts for face images typically include different types of eyes, mouths, etc. Image patches, or *fragments*, are used to depict the appearance of each object part. Fragments are extracted from example images in the learning stage. Each fragment is searched for in the images using a similarity measure, typically the absolute value of normalized cross-correlation, given by

$$NCC(p,F) = \frac{\frac{1}{N}\sum_{x,y}(p(x,y)-\overline{p})(F(x,y)-\overline{F})}{\sigma_p \sigma_F}. \quad (1)$$

Here $F(x,y)$ is the fragment, $p(x,y)$ is an image patch of the same size as F, N is the number of pixels in the fragment, $\overline{p}, \overline{F}$ are the means and σ_p, σ_F are the standard deviations of the intensities of p and F. Image patches at all relevant locations are compared with F, and location with the highest correlation is selected. When the correlation exceeds a pre-determined threshold, the fragment is considered present, or *active*, in the image. An object is represented by the set of fragments that are present in it.

The simplest way to utilize this representation for part matching between two images of the same object is to directly match a fragment from one image with the other image, by normalized correlation. However, this method performs poorly when significant variations in viewing conditions are present [10].

To obtain a reliable match between the same object part under different viewing conditions, consider two fragments, F and F', depicting the same object part P in different viewing conditions C and C'. The key observation is that the part P itself does not change during the transformation, although its appearance changes from F to F'. Therefore, the fragment F, when used with images taken under conditions C, plays an equivalent role to that played by F' in conditions C'. For example, if F is active in the image of some object under conditions C, then F' will be active in the image of the same object under conditions C'. In other words, F and F' will be consistently detected in the images of the same objects, F in conditions C and F' in conditions C'.

Given an arbitrary pair F, F' of fragments, this observation may be used to test whether F matches F'. As discussed above, matching fragments will be consistent. Conversely, non-matching fragments will in general be significantly less consistent. This is because non-matching fragments represent different object parts. In general, different object parts are not highly correlated in different images. Therefore, presence of one fragment will not reliably predict presence of the other, i.e. the fragments will have a low consistency.

More precisely, the consistency of two fragments F, F' may be measured as follows. Assume that a set of images $I_1, \ldots, I_n$ of n objects taken under conditions C, and a set of images $I'_1, \ldots, I'_n$ of the same objects taken under conditions C' are given. (These sets will be called 'validation database'.) For the fragment F, set $A_k = 1$ if F is present in I_k and $A_k = 0$ otherwise. Combine these values into an n-dimensional activation vector A. A is a binary vector with 1's encoding the indices of the objects in which F is present. Similarly, calculate the activation vector A' for F' by $A'_k = 1$ if F' is present in I'_k and $A'_k = 0$ otherwise. Matching fragments will have similar activation vectors. This similarity can be measured by normalized correlation of the vectors A and A', using one-dimensional analogue of eq. (1). Since normalized correlation ranges from -1 to 1, the value

$$C(F,F') = \frac{NCC(A,A')+1}{2} \quad (2)$$

can be used as the consistency measure. $C(F,F')$ ranges from 0 to 1, with 1 indicating perfect consistency and 0 indicating complete inconsistency.

The proposed class-based matching algorithm can now be described as follows. Given two images, I (called 'source image') and I' (called 'target image'), of the same object taken in conditions C, C', the task is to find a dense set of correspondences between I and I'. For every location in I, consider a small image patch F (called 'source fragment') at that location that depicts some object part P under C. In order to find the matching fragment F', consider all possible fragments in I' (called 'candidate target fragments') and select the most consistent fragment F' as the match. Examples of matches obtained by this algorithm are shown in Figure 2 (bottom row).

3.2. Using fragments pyramids to improve accuracy

The previous section explained how fragment consistency is used to evaluate the likelihood of an individual match. However, simply matching the two most consistent features is not the optimal strategy, because factors such as image noise and within-object redundancy may cause matching errors (see Figure 2). A common strategy is therefore to employ some geometric constraints between features to improve the matching accuracy of individual features.

Many existing schemes assume some parametric model (e.g. a homography) of the global scene transformation and derive constraints from this assumption [2, 3, 13]. However, such an assumption is often too restrictive in practice, and therefore more general geometric constraints are necessary. The constraint incorporated in the equivalence-based matching scheme is a simple proximity assumption. Intuitively, we assume that if two object parts are located close to each other in one image, they are likely to remain close in other images.

To impose the proximity constraint, we use a hierarchical representation of the proximity relations between ob-

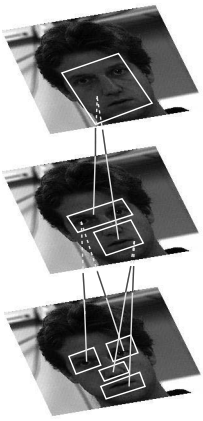


Figure 3: An object is covered by fragments of progressively larger size, depicted by white rectangles. Smaller fragments that fall inside larger ones are considered to be their children, as indicated by gray lines. The resulting hierarchical structure is called the fragments pyramid.

ject parts. In this representation, the image is covered by progressively larger fragments, until a single fragment covers the entire image. A smaller fragment, whose central point is inside the area covered by a larger fragment, is considered a child of the larger fragment in the hierarchy. In order to ensure that each small fragment has exactly one parent, the large fragments are selected to *tile* the image, i.e. to cover the entire image area and be non-intersecting. Since by construction the smallest fragments have no children, the tiling requirement is not applicable at this level. Therefore, fragments of the smallest size are created at every image location. In this hierarchy, two fragments that are spatially close, will usually have a common parent, while two distant fragments are likely to have different parents. In this manner, proximity will be represented by the hierarchicy, as illustrated in Figure 3. The resulting structure is called the fragments pyramid. Since proximity relations are roughly preserved across the transformation of viewing conditions, the hierarchical structure in the source and target images should be similar. Deviations from this common structure can therefore be used as indications of unlikely matches. The next section describes how the fragments pyramid is used to impose the proximity constraint, and how it is combined with the consistency measure to produce the final matching.

3.3. Combining consistency and proximity constraints

The final matching strategy combines the two factors described so far, the likelihood of individual matches (measured by fragment consistency) and the similarity of geometrical structure (measured by fragments pyramid). We describe below how this combination is performed in a probabilistic framework.

First, a fragments pyramid is constructed from the source image. The process is started by creating fragments of certain initial size at every image location. The choice of the initial size may affect the accuracy of the matching, as illustrated by the following example. Consider two fragments F and F' of size one pixel each. Since a single-pixel fragment can be detected in any image, every element of the activation vectors of both F and F' will be equal to one. The two fragments will then be entirely consistent regardless of whether they actually match. To avoid this problem, fragmens should be of appropriate size so that they represent meaningful object parts. The optimal size is selected automatically by testing fragments of several sizes at every image location and selecting the fragment with maximal mutual information, as suggested in [7].

The maximal size of the fragments thus created is then found and increased by a certain factor s. (In our experiments, the best value found empirically was $s = 1.5$.) Larger fragments that tile the source image are created, and parent-child relationships are determined. The process of increasing fragment size is repeated until a single fragment covers the entire image. This fragment is considered the root node of the hierarchy.

Next, candidate target fragments of all sizes used in the source image are created in the target image. Note that the fragment pyramid is constructed for the source image only. Therefore, candidate target fragments are created at every position in the target image, and the constraints of tiling or mutual information are not enforced.

The problem we now face is to establish matches between fragments in the source and target images. Denote by X the unknown vector of matches, where $X_f = f'$ if fragment f matches f'. (The unprimed variables below refer to source fragments and the primed variables refer to target fragments.) Denote by Y the set of observations, which in our case include the consistency values $C(f, f')$ for every pair of fragments f, f', with f taken from the source image and f' from the target image. The commonly used maximum a posteriory (MAP) estimate of the vector of matches is given by

$$\hat{X} = \arg\max_X P(X, Y), \tag{3}$$

where $P(X, Y)$ is the joint probability distribution of X and Y. As shown in [17], under reasonable assumptions (such as conditional independence) $P(X, Y)$ can be factored as

$$\begin{aligned} P(X,Y) \quad = \quad & P(R \rightsquigarrow R') \prod P(f \rightsquigarrow f' | \pi(f) \rightsquigarrow f'_\pi) \times \\ & \prod P(C(f,f') | f \rightsquigarrow f'). \end{aligned} \tag{4}$$

The symbol $f \rightsquigarrow f'$ means that f matches f', $\pi(f)$ denotes the parent of the fragment f in the image pyramid, f'_π is

some fragment from the target image of size similar to $\pi(f)$, and R denotes the root node of the pyramid.

By Bayes formula,

$$P(C(f,f')|f \rightsquigarrow f') = P(f \rightsquigarrow f'|C(f,f'))\frac{P(C(f,f'))}{P(f \rightsquigarrow f')}. \quad (5)$$

Since high consistency is an evidence of a match, it is reasonable to assume that $P(f \rightsquigarrow f'|C(f,f'))$ is proportional to $C(f,f')$. Assuming also that

$$\frac{P(C(f,f'))}{P(f \rightsquigarrow f')} = const, \quad (6)$$

we arrive at the following estimate:

$$P(C(f,f')|f \rightsquigarrow f') = \frac{1}{Z_1}C(f,f'), \quad (7)$$

where Z_1 is an appropriate normalization factor. In principle, the actual distribution $P(C(f,f'))$ could be estimated from examples instead of using (6), but good performance was obtained without doing so. Experiments with a variety of other estimates showed that the algorithm is not sensitive to the precise form of $P(C(f,f')|f \rightsquigarrow f')$.

$P(f \rightsquigarrow f'|\pi(f) \rightsquigarrow f'_\pi)$ implements the proximity constraint, using the hierarchical structure. Assume that $\pi(f) \rightsquigarrow f'_\pi$. By definition, f is a child of $\pi(f)$. Similarity of the hierarchical structure requires f' to be a child of f'_π. Therefore, $P(f \rightsquigarrow f'|\pi(f) \rightsquigarrow f'_\pi)$ should be high if the smaller fragment f' falls inside the larger fragment f'_π, and should decrease when f' becomes more distant from f'_π. The actual distribution $P(f \rightsquigarrow f'|\pi(f) \rightsquigarrow f'_\pi)$ could be learned from examples. However, since no observable examples of correct matches are available in our setting, the following estimate conforming to the qualitative requirements listed above was used:

$$P(f \rightsquigarrow f'|\pi(f) \rightsquigarrow f'_\pi) = \frac{1}{Z_2}\exp(-d). \quad (8)$$

Here d is the distance from the center of f' to the closest point of f'_π, and Z_2 is an appropriate normalization factor. In our experiments, this estimate was sufficient to obtain good performance. Experiments with a variety of other estimates showed that the algorithm is not sensitive to the precise form of $P(f \rightsquigarrow f'|\pi(f) \rightsquigarrow f'_\pi)$.

The root node R in our setting represents the entire image. Since the images to be matched are usually of similar size, there is only one possible match for R. Therefore, $P(R \rightsquigarrow R')$ was set to 1.

Given the decomposition (4) and the values of the factors (7), (8), the maximum a posteriori estimate of each fragment's best match can be calculated efficiently using a standard Viterbi-like inference algorithm [17].

4. Experimental comparisons

In this section, we compare the accuracy of the proposed fragment equivalence scheme to several well-known matching schemes, namely, KLT [1], Black's robust optical flow [9], and affine-invariant features [3]. KLT implementation available at [18], Black's original implementation of robust optical flow available at [19], and Mikolajczyk's original implementation of invariant features available at [20] were used for the experiments.

KLT and robust optical flow were applied to each image pair independently. For the affine invariants scheme, invariant points for each image were calculated. Region matching was then performed by selecting for each source point the target point with the most similar descriptor. Similarity of descriptors was measured by the Mahalanobis distance, using the covariance matrix estimated during training stage from all the images in the database (a separate matrix was used for each experiment below and each database). This evaluation is identical to the published description of the algorithm [3]. The current fragment equivalence scheme was evaluated by using as validation database the entire image set excluding the two images being matched. In addition, the accuracy of the proposed method with and without using fragments pyramids was compared. This comparison demonstrated that the fragments pyramid indeed improves the matching accuracy. In addition, it shows that even without using pyramids, class-based knowledge significantly improves the matching over the previous generic matching schemes.

The data sets used for each of the experiments are described below. The performance of the evaluated algorithms is summarized in Table 1, and the complete histograms of error distributions are shown in Figure 4.

Illumination data set – easy A subset of 30 frontal face images from the PIE database [21] was used in this experiment. Images were taken with normal room illumination. In addition, in the source images a flash from the far right direction was added, and in the target images a flash from the far left. The images were low-pass filtered and downsampled to size 243×320 pixels. Examples are shown in Figure 2 (top row).

Illumination data set – hard Images were similar to the previous test, with room illumination off. As shown in Figure 2 (top row), the changes introduced by illumination variations are much more severe. In particular, illumination can no longer be approximated by a local affine transformation of intensities.

Pose data set A subset of 50 face images from the FERET database [22] was used in this experiment. Frontal images

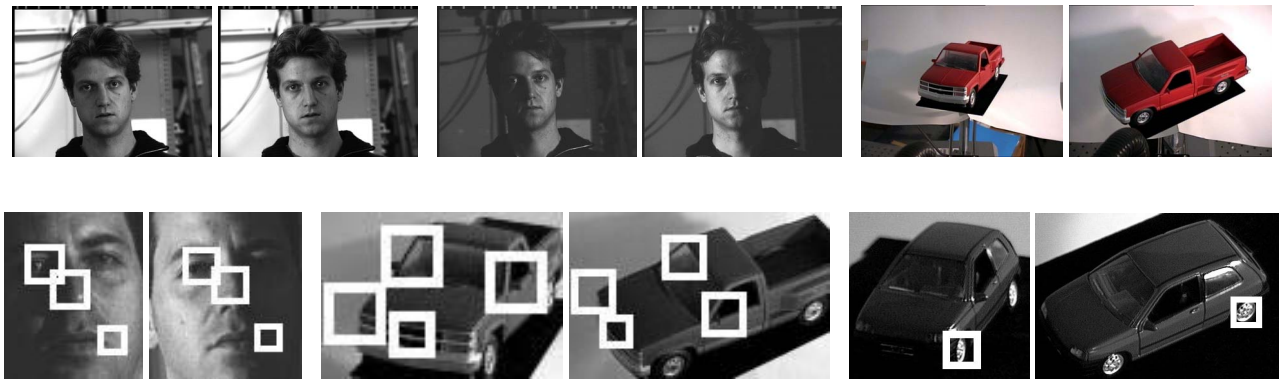
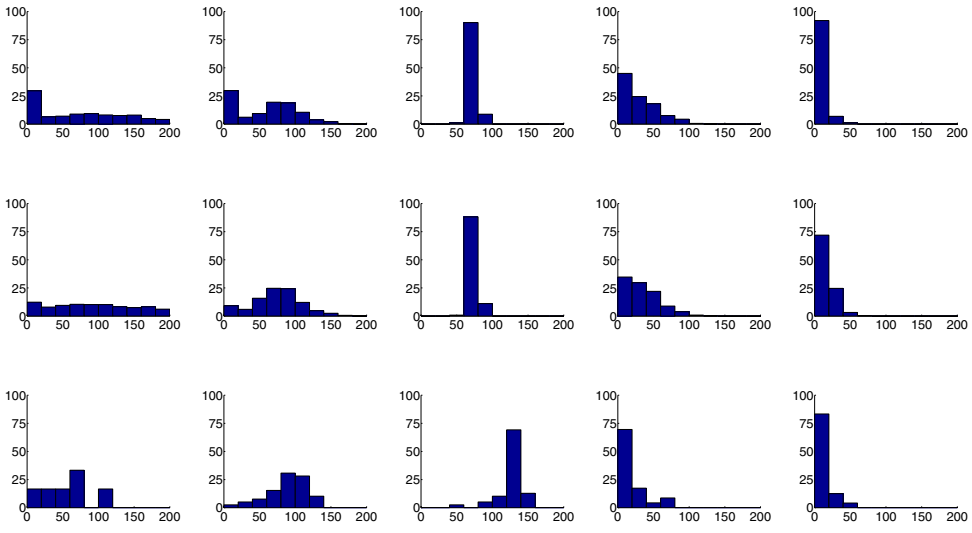


Figure 2: Top row: sample images used for the experiments. Left: easy illumination data set. Middle: hard illumination data set. Right: cars data set. (See sections 4, 5.) Bottom row: examples of matches obtained automatically by the algorithm. Only a few matches are shown, the total number of matches was above 100 in each image. Note that object parts are matched accurately despite significant changes in illumination (including cast shadows) and pose. The rightmost example shows an incorrect match caused by within-object redundancy. We explain in section 3.2 how such errors are handled.

Algorithm	Easy illumination	Hard Illumination	Pose
Affine invariants	82 ± 71	105 ± 69	56 ± 34
KLT	58 ± 43	74 ± 34	89 ± 29
Robust optical flow	73 ± 5	74 ± 5	125 ± 20
Equivalence, no pyramid	29 ± 24	34 ± 23	19 ± 19
Equivalence, with pyramid	8.5 ± 8	15 ± 10	12 ± 11.5

Table 1: Average errors in matches $\pm$ standard deviation, in pixels.



(a) Affine invariants (b) KLT (c) Robust optical flow (d) Equivalence, no pyramid (e) Equivalence with pyramid

Figure 4: Histograms of accuracy of matches. Horizontal axis: error (in pixels). Vertical axis: percentage of points with the given error. Top row: easy illumination data set. Middle row: hard illumination data set. Bottom row: pose data set.

were used as source, and half-profile images were used as target. The images were of size 384×256 pixels.

4.1. Summary

As seen from the results presented above, algorithms that utilize variations of the brightness constancy assumption (KLT and optical flow) perform relatively poorly in all tasks. This is due to significant changes in appearance between the source and target images. In addition, both algorithms assume small feature displacements between images and are usually used with video sequences. Affine invariants perform reasonably well in the easy illumination task (as seen in Figure 4, about 30% of the matches are correct). However, in the more difficult illumination task and in pose task, the performance of affine invariants reduces to random. This is due to the fact that image changes induced by cast shadows and pose change cannot be well approximated by local affine models of illumination and geometry. The performance of the current fragment equivalence scheme remains good in all tasks (above 70% correct matches, Figure 4), although the two latter tasks are more difficult for this scheme as well. The results demonstrate that matching across significant changes in viewing conditions can be achieved by using the appropriate class-based information.

5. Application to invariant recognition

In this section, we illustrate an application of the matching approach presented above to view-invariant object recognition. In the experiments, the system is presented with a single picture of an object (e.g. face of a particular individual or a specific model of a car), taken under certain viewing conditions. The task is then to identify other images of the same object in arbitrary viewing conditions (i.e. other images of the same person or the same car model). The scheme that was used for recognition is an extended version of [6,7], described in [8]. Briefly, an object from a general class is represented by the set of object parts, as described in section 3.1. A set of fragments is used to depict each part under all relevant viewing conditions. This set is called an *extended fragment*. The extended fragment is said to be present in an image if one of its constituent fragments is present. Since each extended fragment contains information regarding the appearance of the given object part in all viewing conditions, its presence or absence in the image depends only on the object and not on the viewing conditions. Therefore, the list of extended fragments that are active in an image forms a view-invariant signature for the object, and this signature may be used for subsequent invariant recognition.

An important step in the scheme is the extraction of extended fragments. This step requires the matching of corresponding object parts across viewing conditions. In [8], video sequences were used to obtain the matches. Such video sequences are not applicable to illumination changes, and are not always available for pose variations. We show that the matches can be obtained by the fragment equivalence scheme described above, without using video sequences and without compromising performance.

Data set	$\lvert\mathcal{T}\rvert$	$\lvert\mathcal{D}\rvert$	Performance (in %)
Illumination – easy	58	10	100
Illumination – hard	58	10	98 ± 4
Pose – faces	40	10	93 ± 5
Pose – cars	28	5	88 ± 14

Table 2: Percentage of correct recognition (average $\pm$ standard deviation). $|\mathcal{T}|$ is the number of training images, $|\mathcal{D}|$ is the number of distractors.

In our experiments, four data sets were used. The pose data set was described above in section 4. The easy illumination and hard illumination data sets were as described in section 4, except that images of 68 individuals were used. In addition, a data set consisting of 33 toy cars viewed from two widely separated directions was used. Examples are shown in Figure 2 (top row).

Each data set was randomly divided into training and testing groups. From the training group, matching fragments were extracted using the fragment equivalence method described above. These matching fragments formed extended fragments, which were used in the recognition stage to represent novel objects of the same class in an invariant manner. During recognition, a single picture of a given object (called 'target object') in certain viewing conditions was presented. The task was to recognize the target object in significantly different viewing conditions among a set of distractors. Images of objects of the same class as the target object were used as distractors.

The performance of the algorithm is summarized in Table 2. The results illustrate that class-based matching by equivalence can be used successfully for the task of invariant recognition.

6. Discussion

A class-based scheme for matching object parts across significant changes in viewing conditions was described. The scheme makes no restrictive assumptions about the transformation, and it performs well under difficult conditions that cannot be handled effectively by previous matching schemes. The scheme is applicable to a variety of natural object classes, it is entirely automatic and does not require examples of correct matches.

In order to perform matching, the scheme uses a validation database of objects of the same class and in similar viewing conditions as the images to be matched. In a series of tests on the pose data set (section 4), the orientation of the

source and target images was changed systematically from the orientation in the validation database. The decrease in performance in these conditions did not reach significance until the difference in orientation was $\pm 25^\circ$. The conclusion is that the conditions in the validation database need not match exactly the conditions of the source and target images.

In assessing the consistency of a possible match, the scheme uses two sets of images, taken under different viewing condition, say conditions C and C'. In practice, images taken under different conditions may be mixed, and it will not be known which images belong to C and which to C'. A simple strategy to deal with this problem is to detect each fragment in every image of the given object, and select the highest score. We have compared this strategy with detecting a fragment only in images under known, correct viewing conditions. The results showed that restricting the detection to the correct viewing conditions does not yield a significant increase in performance. The explanation is that, due to the fact that viewing conditions change significantly the appearance of object parts, fragments are automatically detected almost exclusively in the correct viewing conditions. The conclusion is that when images are not labeled by viewing conditions, fragments can still be extracted correctly by detecting them in all images, without compromising performance. All the results presented above have been obtained using this strategy.

An application to the problem of invariant object recognition was presented. An additional possible application is to use class-based information in a similar manner to improve current stereo techniques. Most current matching techniques for wide-baseline stereo rely on affine-invariant features. The accuracy of matches obtained by these features is reduced significantly when non-affine transformations are present, for example, due to large pose changes of non-planar regions, the effects of highlights, shadows, etc. For highly familiar objects, such as faces, class-based matching techniques could be used to improve the accuracy of matches. The suggested approach is to use the fragment equivalence scheme described above, to obtain matches between object parts as the first stage. Then affine invariants techniques [2–4] may be applied locally, within small regions of the matched fragments, to refine the correspondences. Since the accuracy of the affine approximation improves for smaller regions, the matches will become more accurate. (Small regions cannot be used globally due to matching ambiguities.) Performing stereo matching and 3D reconstruction in this manner is the subject of future work.

Acknowledgements

This research was supported in part by the Moross Laboratory at the Weizmann Institute of Science. Portions of the research in this paper use the FERET database of facial images collected under the FERET program [22].

References

[1] C. Tomasi and T. Kanade, "Detection and tracking of point features," Carnegie Mellon University, Tech. Rep. CMU-CS-91-132, Apr. 1991.

[2] T. Tuytelaars and L. V. Gool, "Wide baseline stereo matching based on local, affinely invariant regions," in *British Machine Vision Conference*, 2000, pp. 412–425.

[3] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *Proceedings of the European Conference on Computer Vision*, 2002, pp. 128–142.

[4] M. Brown and D. Lowe, "Invariant features from interest point groups," in *British Machine Vision Conference*, 2002.

[5] G. D. Hager and P. N. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 10, pp. 1025–1039, 1998.

[6] S. Agarwal and D. Roth, "Learning a sparse representation for object detection," in *Proceedings of the European Conference on Computer Vision*, 2002, pp. 113–127.

[7] S. Ullman, M. Vidal-Naquet, and E. Sali, "Visual features of intermediate complexity and their use in classification," *Nature Neuroscience*, vol. 5, no. 7, pp. 682–687, 2002.

[8] E. Bart, E. Byvatov, and S. Ullman, "View-invariant recognition using corresponding object fragments," in *Proceedings of the European Conference on Computer Vision*, 2004, to appear.

[9] M. J. Black and P. Anandan, "A framework for the robust estimation of optical flow," in *Proceedings of the International Conference on Computer Vision*, 1993, pp. 231–236.

[10] Y. Adini, Y. Moses, and S. Ullman, "Face recognition: the problem of compensating for changes in illumination direction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 721–732, 1997.

[11] S.-H. Lai, "Robust image matching under partial occlusion and spatially varying illumination change,"

Computer Vision and Image Understanding, vol. 78, pp. 84–98, 2000.

[12] M. J. Black and A. D. Jepson, "EigenTracking: Robust matching and tracking of articulated objects using a view-based representation," in *Proceedings of the European Conference on Computer Vision*, 1996, pp. 329–342.

[13] D. Tell and S. Carlsson, "Combining appearance and topology for wide baseline matching," in *Proceedings of the European Conference on Computer Vision*, 2002, pp. 68–81.

[14] D. G. Lowe, "Three-dimensional object recognition from single two-dimensional images," *Artificial Intelligence*, vol. 31, no. 3, pp. 355–395, 1987.

[15] P. Viola and W. Wells, "Alignment by maximization of mutual information," *International Journal of Computer Vision*, vol. 24, no. 2, pp. 137–154, 1997.

[16] A. Chowdhury, R. Chellappa, and T. Keaton, "Wide baseline image registration with application to 3D face modeling," accepted to IEEE Trans. Multimedia.

[17] J.-M. Laferte, P. Perez, and F. Heitz, "Discrete markov image modeling and inference on the quadtree," *IEEE Transactions on Image Processing*, vol. 9, no. 3, pp. 390–404, 2000.

[18] http://vision.stanford.edu/ birch/klt/.

[19] http://www.cs.brown.edu/people/black/ignc.html.

[20] http://www.inrialpes.fr/lear/people/Mikolajczyk/.

[21] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression (PIE) database of human faces," The Robotics Institute, Carnegie Mellon University, Tech. Rep. CMU-RI-TR-01-02, January 2001.

[22] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The FERET database and evaluation procedure for face recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295–306, 1998.