

Identifying Semantically Equivalent Object Fragments

Boris Epshtein

Department of Computer Science and Applied Mathematics

Weizmann Institute of Science

Rehovot, ISRAEL, 71600

{boris.epshtein , shimon.ullman}@weizmann.ac.il

Shimon Ullman

Abstract

We describe a novel technique for identifying semantically equivalent parts in images belonging to the same object class, (e.g. eyes, license plates, aircraft wings etc.). The visual appearance of such object parts can differ substantially, and therefore traditional image similarity-based methods are inappropriate for this task. The technique we propose is based on the use of common context. We first retrieve context fragments, which consistently appear together with a given input fragment in a stable geometric relation. We then use the context fragments in new images to infer the most likely position of equivalent parts. Given a set of image examples of objects in a class, the method can automatically learn the part structure of the domain – identify the main parts, and how their appearance changes across objects in the class. Two applications of the proposed algorithm are shown: the detection and identification of object parts and object recognition.

1. Introduction

In this paper we consider the problem of detecting semantically equivalent parts of objects belonging to the same class. By ‘semantic’ equivalence we mean parts of the same type in similar objects, often having the same part name, such as an eye in a face, an animal’s leg, an airplane’s wing and the like. The goal is to identify such parts even when their visual appearance is highly dissimilar. The input to the algorithm is a set of images belonging to the same object class, together with an image patch (called below a “root fragment”), depicting a part of an object. The output is a set of image patches from the input images, containing object parts which are semantically equivalent to the part depicted in the root fragment. For instance, taking as input a set of face images, and a root fragment containing a nose, the algorithm identifies image regions containing noses in other input images. Examples are shown in Figure 1. In

each row, the leftmost image contains the root fragment, the other images are equivalent fragments discovered by the algorithm.

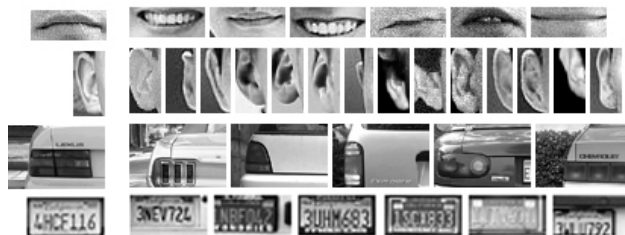


Figure 1: Examples of semantically equivalent fragments, extracted by the algorithm. The leftmost image in each row is the input root fragment, the others are equivalent parts identified by the algorithm (mouth, ear, tail-light, license plate). The algorithm identifies semantically similar parts that can have markedly different appearance in the image.

The identification of equivalent object parts has two main goals. First, the correct detection and identification of object parts is important on its own right, and can be useful for various applications: recognizing facial expressions, visual aid for speech recognition, visual inspection, surveillance and so on. Second, as shown in Section 4.2, the correct identification of semantically equivalent object parts improves the performance of object recognition algorithms. In several recent object recognition schemes [11-15], image fragments depicting object components are used as classification features. Our results show that the performance of such schemes can be improved when object components are represented not by a single fragment, but by a set of semantically equivalent fragments.

The general idea behind our approach is to use common context to identify equivalent parts. Given an image fragment F depicting a part of an object, we look for a context C , defined as a collection of image fragments that co-occur with F consistently and in a stable geometric configuration. When such context is found, we look for all images where the context fragments are

detected, and infer from their positions the location of fragments that are likely to be semantically equivalent to F .

The rest of the paper is organized as follows. In the next section we briefly review previous approaches to the problem of identifying equivalent object parts. In Section 3 we describe the proposed algorithm. In Section 4 we show experimental results, comparing the use of semantically equivalent parts with visually similar parts in two tasks: identifying object parts, and object detection. We conclude with general remarks on possible extensions and applicability of the method in Section 5.

2. Previous work

In most object recognition schemes, similar object components are identified by their image-based similarity, using similarity measures such as L_2 norm, or normalized cross-correlation. Simple image-based similarity measures were found in practice to be too restrictive, since they are sensitive to the viewing conditions, such as changes in illumination and viewing direction. More flexible similarity measures, such as the SIFT measure [1] proved more successful, since they rely on the general configuration of image gradients rather than the exact grey level distribution within an image patch. Another approach to deal with shape variability of local object components has been to use affine-invariant matching [2-4]. Such methods can efficiently identify local components that differ by an affine transformation, and can be useful for identifying the same image part at different scales and viewed from different directions. However, our task requires the identification of similar parts that differ by more than an affine transformation, such as an open or closed eye, neutral or smiling mouth, different hairlines, and the like.

Another approach to the problem of generalization of object parts is “extended fragments”, which are the equivalence classes of image fragments, representing object parts under different viewing conditions [5]. This method identifies corresponding object parts using motion tracking, and it will not be applicable when motion sequences are not available, or for identifying equivalent parts in images of different objects.

The role of context in object recognition has been studied in [6-9]. In these studies, the context is represented either by global statistical properties of the scene [6-7] or by classification labels of co-occurring objects [9]. Context is used to analyze objects within a scene, rather than individual objects. Our method retrieves context as a collection of patches, co-occurring within an object in a stable geometric configuration.

3. Description of the algorithm

In this section, we describe the algorithm for the detection of semantically equivalent image fragments. The main stages of the algorithm are the identification of common context (3.2) and the use of context to extract equivalent parts (3.3). We begin with describing visual similarity matching used as a pre-processing step.

3.1. Visual similarity matching

The input to the algorithm consists of a set of images of different objects within a class, I_k , and a single fixed fragment F (the “root fragment”). We first identify in each of the input images I_k the image patch with the maximal similarity to F . We used the value of normalized cross-correlation as a similarity measure (other image-based similarity measure can be used, see Section 5). To improve the performance of visual similarity-based matching the images were filtered with Difference of Gaussians (DoG) filter [1] before computing the NCC. This filter emphasizes the gradients in images and removes small noise. The combination of DoG filtering with computation of NCC is called below DNCC.

Image patches at all the locations in I_k are examined, and the patch $P(I_k, F)$ with highest DNCC score is selected. If the cross-correlation between $P(I_k, F)$ and F exceeds a pre-defined threshold, then F is detected in I_k , and we call $P(I_k, F)$ the patch corresponding to F in image I_k . The set of all the images I_k where corresponding patches $P(I_k, F)$ were detected is denoted by $D(F)$. The detection threshold for candidate context patches was chosen automatically as in [11]. Briefly, the detection threshold of a patch was chosen to maximize the mutual information between the class variable (1 in images containing object, 0 otherwise) and patch variable (1 when the patch was detected in image, 0 otherwise).

3.2. Context retrieval

After determining the set $D(F)$, containing the images where F was detected, the next goal is to identify context fragments that consistently co-occur with F and its corresponding patches $P(I_k, F)$. Reliable context fragments should meet two criteria: the context fragment f and root fragment F should have high probability of co-occurrence, and their spatial relations should be stable. We next describe the selection based on these criteria.

The search starts by pairing the root F with patches f_i in each image in $D(F)$ at multiple sizes and positions. These patches are the candidate context patches for F . We used patch sizes ranging from 50% of F size up to 150% in each dimension, with scaling step of 1.5. For each patch size, we examine patches in positions placed on a

regular grid with step equal to $\frac{1}{4}$ of the size of a patch. The exact position and size of a patch is optimized as described later in this Section. For every candidate patch f , we find the set $D(f)$ of images containing patches visually similar to f , as described in Section 3.1.

The first context condition above was high co-occurrence, that is, a good context fragments should satisfy $p(F|f) > p(F)$. We also want to focus on context fragments that appear together with F at least some minimal number of times:

$$P(f|F) > \theta_p \quad p(F|f) > p(F) \quad (1)$$

The value of θ_p was computed automatically by sampling a set of candidate patches from $D(f)$, computing their probabilities of co-occurrence with F and setting the threshold to average co-occurrence probability plus a standard deviation.

Second, F and f should appear in a stable spatial configuration. If the variations in scale and orientation between the images are assumed to be small (we relax these assumptions in Section 5), then the relative location of F and f when they are detected together, should be similar. We therefore test the variance of coordinate differences:

$$\text{Var}(F_x - f_x) < \theta_{\text{var}X} \quad \text{Var}(F_y - f_y) < \theta_{\text{var}Y} \quad (2)$$

Here F_x and f_x are vectors of x -coordinates of the centers of image patches corresponding to F and f , respectively, in images from $D(F) \cap D(f)$, similarly for F_y and f_y . The thresholds $\theta_{\text{var}X}$ and $\theta_{\text{var}Y}$ determine the flexibility of the geometric model of the object. These thresholds were also set automatically by computing the values of $\text{Var}(F_x - f_x)$ and $\text{Var}(F_y - f_y)$ for the sampled fragments, for which $P(f|F) > \theta_p$ and setting the thresholds to the average of these values plus a standard deviation.

To identify the best context fragments, we first remove from the set of candidates all fragments that do not meet requirements (1) and (2). We next select from the remaining set the fragments with the highest probability of co-occurrence with F , and smallest variances of coordinate differences (indicating a stable geometric relation with the root F). To combine these criteria, we compute a ‘consistency weight’ w_f .

$$w_f = P(f|F) \cdot \frac{1}{1 + \sqrt{\max(\text{Var}(F_x - f_x), \text{Var}(F_y - f_y))}} \quad (3)$$

The fragment with the highest w_f is then selected as a context fragment. Since the initial search for context fragments was limited to a fixed grid, we refine the optimal position and size of the context fragment by searching locally for the best fragment position and size

that maximize w_f . We add the optimized fragment to the set of context fragments. To avoid redundancy, and prefer conditionally independent context fragments (see Section 3.3 for details), we remove from the set of remaining candidates all the fragments that intersect the selected one by more than 25% of their area, and repeat the process until no candidates are left. The final context set contains fragments f_i that have high co-occurrence with F , and with stable relative positions. Typically this set contains between 6 and 12 fragments.

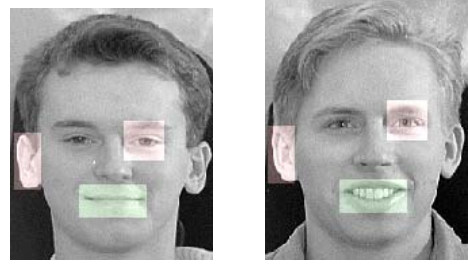


Figure 2: Left: a root fragment (mouth) together with context fragments (ear and eye). Right: the same context detected in another image; a semantically equivalent part is identified by the context.

3.3. Identifying semantically equivalent parts

After the set of context fragments has been selected, they are used to infer the positions of fragments that are semantically equivalent to the root fragment F . Using a probabilistic model, we identify for each image I_k , in which at least one context fragment has been detected, the most likely position of F_k , a semantically equivalent fragment to F .

Assume first for simplicity that our context set consists of a single fragment C . Our modeling assumption is that if C is detected in some image I_k at coordinates (x_c, y_c) , then the probability density of F being found at coordinates (x, y) is a 2D Gaussian centered at (\hat{x}_c, \hat{y}_c) , where \hat{x}_c and \hat{y}_c are the expected coordinates of the root fragment center, predicted by context fragment C . The values of \hat{x}_c and \hat{y}_c are computed as:

$$\hat{x}_c = x_c + \overline{\Delta x_c} \quad \hat{y}_c = y_c + \overline{\Delta y_c} \quad (4)$$

where $\overline{\Delta x_c}$ and $\overline{\Delta y_c}$ are the mean coordinate differences between the centers of F and C , estimated during training.

$$P(F(x, y) | C) = P(F | C) \cdot N(x - \hat{x}_c, y - \hat{y}_c; \Sigma_c) \quad (5)$$

where Σ_c is the covariance matrix of coordinate differences between the centers of fragments F and C , estimated during training.

If the context fragment C is not detected in the image I_k , we assume 2D uniform probability density of F being found at coordinates (x, y) :

$$P(F(x, y) | \bar{C}) = P(F | \bar{C}) \cdot U(W, H) \quad (6)$$

here the distribution bounds W and H are set to the width and height of the image.

When the context consists of several fragments, we assume conditional independence between them given the detection of F at position (x, y) :

$$P(C_1, \dots, C_N | F(x, y)) = \prod_{i=1}^N P(C_i | F(x, y)) \quad (7)$$

The modeling assumption of the conditional independence is motivated by the observation that if the geometric relation between fragments is stable, the positions of the context fragments are determined by the position of the root fragment. The fluctuations of the positions are due to noise, which is assumed to be independent for the context fragments. Modeling of higher-order geometric relations between fragments is also possible, but we found in testing that it did not make a significant contribution. Applying Bayes rule to (7):

$$\begin{aligned} P(F(x, y) | C_1, \dots, C_N) &= \\ &= \frac{P(F(x, y))}{P(C_1, \dots, C_N)} \prod_{i=1}^N P(C_i | F(x, y)) \end{aligned} \quad (8)$$

We assume the prior probability $P(F(x, y))$ of finding F at the coordinates (x, y) to be uniform, therefore not depending on x and y . It is also straightforward to use a non-uniform prior (see Section 5). The probability $P(C_1, \dots, C_N)$ similarly does not depend on (x, y) . Therefore, we can write:

$$P(F(x, y) | C_1, \dots, C_N) \propto \prod_{i=1}^N P(C_i | F(x, y)) \quad (9)$$

For the individual factors $P(C_i | F(x, y))$ we use equations (5) or (6), depending on whether or not the context fragment C_i was detected in the image. Again, we apply Bayes rule:

If C_i was detected in the image:

$$\begin{aligned} P(C_i | F(x, y)) &= \frac{P(C_i) \cdot P(F(x, y) | C_i)}{P(F(x, y))} = \\ &= \frac{P(C_i) \cdot P(F | C_i) \cdot N(x - \hat{x}_{ci}, y - \hat{y}_{ci}; \Sigma_{Ci})}{P(F(x, y))} \end{aligned} \quad (10)$$

If C_i was not detected in the image:

$$\begin{aligned} P(\bar{C}_i | F(x, y)) &= \frac{(1 - P(C_i)) \cdot P(F(x, y) | \bar{C}_i)}{P(F(x, y))} = \\ &= \frac{(1 - P(C_i)) \cdot P(F | \bar{C}_i) \cdot U(W, H)}{P(F(x, y))} \end{aligned} \quad (11)$$

We can now find the values of coordinates (x, y) that maximize (9), i.e. find a Maximum Likelihood solution for the coordinates of the center of the fragment F :

$$(x, y) = \arg \max \prod_i N(x - \hat{x}_{ci}, y - \hat{y}_{ci}; \Sigma_{Ci}) \quad (12)$$

where each 2D Gaussian can be explicitly written in terms of its parameters: mean position and covariance matrix. Taking the log of the product, differentiating with respect to x and y , and setting the derivatives to zero, yields a system of equation of the form:

$$xA - yB + C = 0 \quad yD - xB + E = 0 \quad (13)$$

where

$$\begin{aligned} A &= \sum_i \frac{1}{(1 - \rho_{xyi}^2) \sigma_{xi}^2} & B &= \sum_i \frac{\rho_{xyi}}{\sigma_{xi} \sigma_{yi}} \\ C &= \sum_i \left(\frac{\rho_{xyi} (y_{ci} + \overline{\Delta y_{ci}})}{\sigma_{xi} \sigma_{yi}} - \frac{x_{ci} + \overline{\Delta x_{ci}}}{(1 - \rho_{xyi}^2) \sigma_{xi}^2} \right) \\ D &= \sum_i \frac{1}{(1 - \rho_{xyi}^2) \sigma_{yi}^2} \\ E &= \sum_i \left(\frac{\rho_{xyi} (x_{ci} + \overline{\Delta x_{ci}})}{\sigma_{xi} \sigma_{yi}} - \frac{y_{ci} + \overline{\Delta y_{ci}}}{(1 - \rho_{xyi}^2) \sigma_{yi}^2} \right) \end{aligned} \quad (14)$$

$$\begin{aligned} \sigma_{xi} &= \sqrt{\text{Var}(x - x_{ci})}, & \sigma_{yi} &= \sqrt{\text{Var}(y - y_{ci})}, \\ \rho_{xyi} &= \frac{\text{Cov}((x - x_{ci}), (y - y_{ci}))}{\sigma_{xi} \sigma_{yi}} \end{aligned}$$

Solving (13), we obtain:

$$y = \frac{AE + BC}{B^2 - AD} \quad x = \frac{By - C}{A} \quad (15)$$

After obtaining the ML solutions for the coordinates (x, y) we extract a fragment centered at (x, y) with size equivalent to the size of F (see Section 5 for discussion), and add it to the set of fragments semantically equivalent to F .

The set of semantically equivalent fragments constructed in this manner is called the equivalence set. We next sort it by measuring the strength of the evidence used to select the fragments. This is obtained by setting the optimal values found for (x, y) into (9) and taking the log. The resulting quantity is equal to the log-likelihood of the optimal solution plus a constant factor. This value is then used to sort the equivalence set: the log-likelihood will be smaller when only a few context fragments were

detected in a particular image, or when their evidence was inconsistent, i.e. they predict different locations of a semantic fragment. The decision regarding the number of fragments from the equivalence set to be used is application-dependent. For the object recognition experiments we used the upper 30% of the sorted equivalence set. For the part detection experiments we used the entire set and counted the number of errors.

The section above described the main computation; its accuracy can be improved by incorporating a number of additional steps. We used in particular simple criteria to reject outliers, based on the fact that they will be detected at highly variable image locations. We therefore computed the average value of coordinate differences between the detected positions of F and f , and removed the farthest outliers, until the variance of coordinate differences is below threshold.

The same procedure for outlier rejection is used when performing the ML estimation, since some of the context fragments can correspond to false detections.

3.4. Brief summary of the algorithm

INPUT:

A set T of images belonging to the same object class.
An image fragment F , depicting a part of an object (“root fragment”).

OUTPUT:

A set S (“equivalence set”) of image fragments semantically equivalent to F .

ALGORITHM:

1. Initialize S and C as the empty sets.
2. Find the set $D(F)$ of images where fragments visually similar to F are detected.
3. Form a set K of candidate context fragments from $D(F)$, of various sizes and positions.
4. For every fragment $f \in K$ check conditions (1) and (2), delete from K fragments that fail to meet them.
5. For every remaining fragment $f \in K$ compute the weight w_f according to (3).
6. Find the fragment f^* from K with the largest w_f .
7. Optimize the position and size of f^* for the best w_f .
8. Add f^* to C .
9. Remove from K all fragments that intersect f^* by more than 25% of their area.
10. If K is non-empty – go to step 6
11. For every fragment C_i from the context set C estimate on the training set $P(C_i)$, $P(F | C_i)$, $P(F | \overline{C_i})$. Estimate the covariance matrix Σ_{C_i} on the set $D(F) \cap D(C_i)$

This completes the selection of context set C . The next step uses C to identify the equivalence set S

12. For every image $I_k \in T$, compute the ML estimation for the coordinates (x, y) of fragment F_k semantically equivalent to the root fragment F , using the context set C , as explained in Section 3.3. Compute the likelihood of F_k . If it is larger than a threshold, add F_k to S .
13. Return S .

4. Experimental results

This section describes the results of applying the algorithm to two recognition tasks. Section 4.1 shows the application of the method to the detection of object parts; Section 4.2 describes the application of the method to full object detection.

4.1. Object parts detection

We selected first for testing 7 root fragments depicting different parts of the human face (left ear, right ear, nose, mouth, left eye, right eye with an eyebrow, chin) and applied the algorithm described in Section 3 to detect semantically equivalent parts in new face images, independently for each root fragment. For comparison, we applied the algorithm for detecting face parts based on their visual similarity to the root fragment, as described in Section 3.1, using the same input image set and root fragments. Visual similarity was computed using two different measures – DNCC and SIFT [1]. We applied both algorithms to a database of 1000 face images (about 150x200 pixels in size, of roughly the same scale and orientation) and counted the number of images where all the parts were detected correctly. We also performed object parts detection on a database of toy cars (560 images, roughly 220x150 pixels in size, collected from the Web) and the CALTECH database of cars (rear view, 126 images, 300x200 pixels, [http://www.vision.caltech.edu/Image_Datasets/cars_marcus/cars_marcus.tar]). Both databases presented large variability of car models (Jeeps, sedans, race cars, etc.). Examples of database images together with identified parts are shown in Figure 3. Figure 4 shows the set of context fragments retrieved by the algorithm for a single root fragment

The numbers of face images where all 7 fragments were detected correctly were as follows: Semantic equivalence: 379; DNCC visual similarity: 5; SIFT visual similarity: 7. As can be seen, the method is successful in recovering a large number of correct part configurations, that cannot be identified by their visual similarity. The percentage of correctly identified matches, verified by humans, for semantic equivalence and DNCC visual

similarity was also computed for each individual part, yielding the results in Table 1. Using the SIFT similarity measure produced similar results to DNCC.













Root fragment	Semantic equivalence	Visual similarity (DNCC)
	94%	33%
	92%	39%
	92%	71%
	89%	20%
	90%	29%
	88%	51%
	84%	26%
	65%	41%
	55%	18%
	64%	25%
	71%	59%
	42%	32%

Table 1. Percentage of correctly identified fragments

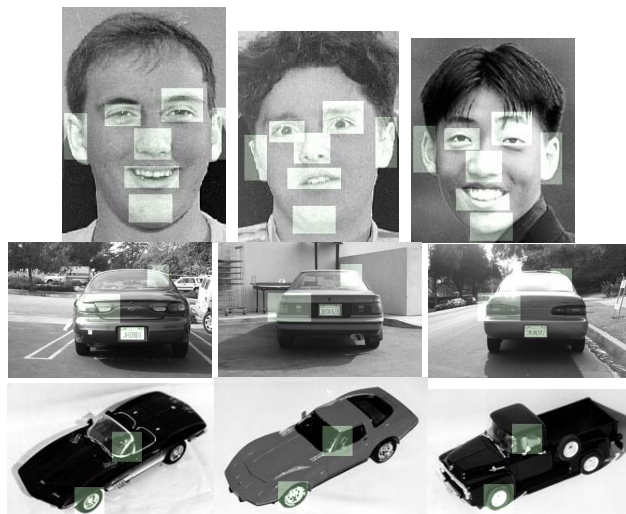


Figure 3: Examples of semantically equivalent fragments overlaid on object images.

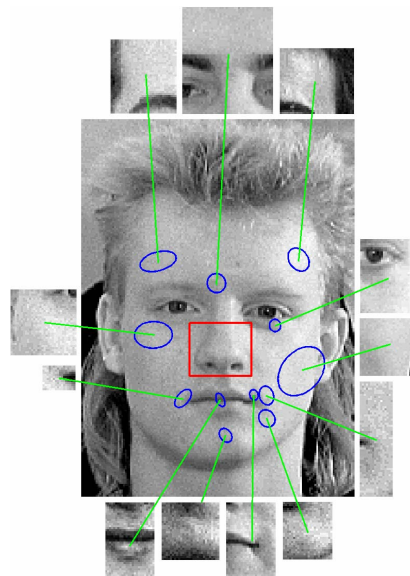


Figure 4: Example of context fragments, retrieved for one root fragment (nose). The sizes of ellipses indicate the variability of geometric relation between context fragment and the root (eq. 2, two sigma).

4.2. Object recognition

The detection and use of equivalent parts is expected to improve object detection, because a face, for instance, is composed of parts such as eyes, hairline etc., which should be detected in the image regardless of their exact appearance. Using our method, equivalent parts can be detected automatically during training and then used for recognition. In this section we show that object recognition results are in fact improved by the use of semantically equivalent parts compared with visually similar object parts. We first describe briefly the classifier used for the comparison, then present the experimental results.

The classifier we used for the experiments is an extension of a classifier described in [11]. Briefly, an object from a general class is represented by a collection of object parts. A set of fragments (visually similar or semantically equivalent) selected automatically is used to represent each part. An object part is detected in an image if one of fragments representing it is detected ($DNCC > \theta_i$, where θ_i is the threshold of fragment i , see Section 3.1.) within a detection window. Each fragment is assigned a weight w_i determined by log-likelihood ratio:

$$w_i = \log \frac{P(F_i | C)}{P(F_i | \bar{C})} \quad (16)$$

where C is a class variable (1 in images containing an object, 0 otherwise) and F_i is a fragment variable (1 when

the fragment was detected, 0 otherwise). Final detection is based on a naïve Bayesian decision,

$$\sum_i w_i F_i > \theta \quad (17)$$

where θ is decision threshold; by varying θ complete ROC curves are obtained (Figure 5).

Detection performance was compared using 7 face parts, as in Section 4.1. Each part was then represented by 20 representative image fragments selected so as to optimize performance. The two schemes we compared used an identical classifier, but differed in the selection of the image fragments representing each part. In the ‘semantic’ scheme, each part was represented by a set of 20 semantically equivalent fragments, selected by the algorithm described in Section 3. In the ‘visual similarity’ scheme each part was represented by 20 representative image fragments, selected from the set of visually similar fragments to the root, so as to optimize performance.

The selection of representative fragments for each face part was done in a greedy fashion, using a mutual information criterion: the fragment delivering the highest information between the classifier response and the true class was selected first. Next, all remaining fragments were examined to find which fragment adds the most information when added to the first one. The process was repeated until 20 fragments have been selected. An identical selection procedure was used to select the best representatives from the set of visually similar fragments.

We divided the image set randomly into a training set (300 faces, 600 non-faces) and test set (700 faces, 1400 non-faces) and repeated the computation 50 times. The results are presented in Fig. 5. Fig. 5a shows the comparison of ROC curves of a single root fragment (the mouth in Table 1): the ROC curve of the classifier based on this fragment alone (red), the ROC curve of the classifier based on visually similar fragments (green) and the ROC curve of the classifier based on semantic equivalence class (blue). Fig. 5b shows the ROC curves of a set of 7 root fragments used for classification together. Fig. 5c shows the mean difference between the ROC curves of a classifier based on visually similar fragments, and the classifier that uses semantic equivalence classes for single parts. Fig. 5d shows the mean difference between the ROC curves of a classifier based on the group of 7 root fragments used together and the group of semantic equivalence classes. Additional comparison was made to a classifier using the same number of fragments ($7 \times 20 = 140$) selected from the set of all the fragments extracted by the algorithm (1400 in total) with no constraints. The performance in this case is worse than the performance of both classifiers based on visually similar, and on semantically equivalent fragments, compared in Fig. 5b. At the equal error point (EEP, misses = false

alarms) semantically equivalent fragments give 1.13% error, visually equivalent 2.89%, 140 best fragments without equivalence 4.44%.

Image similarity was based in the scheme above on normalized cross-correlation. Other, more robust image comparison measures have been introduced recently, which compensate for scale changes, affine transformations, and small local distortions (see [2] for a review). Comparisons in [10] have shown that in the absence of scale changes and affine transformations, the performance of normalized cross-correlation is comparable to the performance of the SIFT descriptor [1] and better than the results obtained by other measures. Since we tested our algorithm under such conditions, the use of DNCC was appropriate. We also compared the performance of DNCC and SIFT, in the following way. For each face image, the semantically equivalent fragment to the root and the visually similar fragment to the root were determined by the algorithm (only the images where both fragments were found by the algorithms were considered). The images where the computed semantic fragment was correct (as determined by an observer), but the fragment selected by visual similarity was incorrect, were chosen. For each image we then normalized the three fragments (the root, the semantically equivalent and the most visually similar) by an affine transform to a normal form [2] and compared the SIFT distance between the root and semantic fragment, to the SIFT distance between the root and the visually similar fragment. In 74.6% of the cases the SIFT made the incorrect selection: the visually similar fragment was closer to the root fragment than the semantic fragment. We conclude that the SIFT distance did not overcome the incorrect choice of the visually similar fragment made by the DNCC.

5. Discussion

The current implementation assumed that the images have roughly similar scale and orientation. This assumption can be relaxed to deal with variable scale and orientation by using context configurations. For example, we can consider pairs of image fragments as elementary units of context. We can then use angles between the line connecting the context fragments and lines connecting them to the root fragment. The semantic fragment is found on the intersection of such lines in the novel image. When rotation or change of scale is detected, we must also rotate and scale the semantic fragment accordingly. Also, the measures for computing visual similarity should be tolerant to changes in scale and rotations (e.g. using SIFT descriptors [1]). The probabilistic model also needs to be modified in a relatively straightforward manner

The method can also be extended by using the semantic fragments themselves as context for other parts.

This can be obtained by an iterative procedure that uses visual similarity to find the first context fragments, but uses semantically equivalent fragments in subsequent iterations. This can be useful in situations when objects appearance is highly variable.

Another extension of the algorithm can employ non-uniform priors on the coordinates of root fragments. This approach is justified when input images are tightly cropped around the objects, or when we have other prior knowledge about the position of the object in the image. Alternatively, the prior probability can be estimated using non-parametric techniques, e.g. 2D Parzen window.

In summary, the work presented a method for identifying semantically equivalent object parts. The scheme makes no restrictive assumptions about visual similarity between semantically identical fragments, and it correctly identifies the same object parts even when they are visually dissimilar. Using this method, equivalent parts can be detected automatically during training and then used for recognition. The method is useful for identifying object parts, and for improving the recognition of complete objects.

Acknowledgements

This research was supported in part by a grant from the Israel Ministry of Science and Technology and by the Moross Laboratory for Vision and Motor Control.

References

- [1] D. Lowe, "Distinctive image features from scale-invariant keypoints", *Int. J. Comp. Vis.* 60(2), pp. 91-100, 2004.
- [2] K. Mikolajczyk and C. Schmid, "Scale & affine invariant point detectors", *Int. J. Comp. Vis.*, 60(1), pp. 63-86. 2004.
- [3] T. Tuytelaars and L. Van Gool, "Wide baseline stereo matching based on local, affinely invariant regions", *Proc. BMVC'00*, pp. 412-425, 2000.
- [4] J. Matas, O. Chum, M. Urban and T. Pajdla, "Robust wide baseline stereo from maximally stable external regions", *Proc. BMVC'02*, pp. 384-393, 2002.
- [5] E. Bart, E. Byvatov, S. Ullman, "View-Invariant Recognition Using Corresponding Object Fragments", *Proc. ECCV*, pp. 152-165, 2004.
- [6] K. Murphy, A. Torralba and W. Freeman. "Using the forest to see the trees: a graphical model relating features, objects and scenes". *Adv. in Neural Information Processing Systems 16* (NIPS), 2003.
- [7] A. Torralba, "Contextual priming for object detection", *Int. J. Comp. Vis.* 53(2), pp. 169-191, 2003.
- [8] T. Strat, "Employing Contextual Information in Computer Vision", *DARPA93*, pp.217-229, 1993.
- [9] X. Song, J. Sill, Y. Abu-Mostafa and H. Kasdan, "Image recognition in context: application to microscopic urinalysis", *Advances in Neural In]ormation Processing Systems*, pp. 963-969, 2000.
- [10] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors", *Proc. CVPR*, pp. 257-264, 2003.
- [11] Ullman, S, Vidal-Naquet, M, Sali, E, "Visaul features of intermediate complexity and their use in classification", *Nature Neuroscience*, 5, pp. 682-687, 2002.
- [12] S. Agarwal and D. Roth, "Learning a sparse representation for object detection", *Proc. ECCV*, pp. 113-127, 2002.
- [13] R. Fergus, P. Perona, A. Zisserman, "Object Class Recognition by Unsupervised Scale-Invariant Learning", *Proc. CVPR* (2), pp. 264-271, 2003.
- [14] B. Leibe and B. Schiele, "Interleaved object categorization and Segmentation", *Proc. BMVC'03*, 2003.
- [15] B. Heisele, T. Serre, M. Pontil, T. Vetter and T. Poggio, "Categorization by learning and combining object parts", *Neural Information Processing Systems*, 2001.

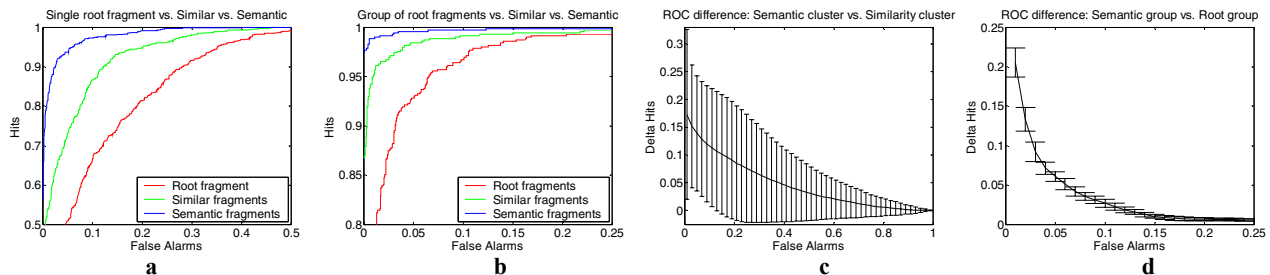


Figure 5: Comparing recognition by semantic and visual similarity. (a) ROC curves for a single part (mouth); (b) classification by 7 parts; (c) average gain in ROC between semantic and visual similarity for individual parts; (d) average gain in ROC for classification by 7 parts using semantic vs. visual similarity.