Robust Detection of Semantically Equivalent Visually Dissimilar Objects

Taeil Goh, Ryan West, Kazunori Okada
San Francisco State University
San Francisco, CA 94132
{imbb,rawest,kazokada}@sfsu.edu

Abstract

We propose a novel and robust detection of semantically equivalent but visually dissimilar object parts with the presence of geometric domain variations. The presented algorithms follow a part-based object learning and recognition framework proposed by Epshtein and Ullman. This approach characterizes the location of a visually dissimilar object (i.e., root fragment) as a function of its relative geometrical configuration to a set of local context patches (i.e., context fragments). This work extends the original detection algorithm for handling more realistic geometric domain variation by using robust candidate generation, exploiting geometric invariances of a pair of similar polygons, as well as SIFT-based context descriptors. An entropic feature selection is also integrated in order to improve its performance. Furthermore, robust voting in a maximum density framework is realized by variable bandwidth mean shift, allowing better root detection performance with the presence of significant errors in detecting corresponding context fragments. We evaluate the proposed solution for the task of detecting various facial parts using FERET database. Our experimental results demonstrate the advantage of our solution by indicating significant improvement of detection performance and robustness over the original system.

1. Introduction

Part-based object recognition [2, 4, 5, 7] is an effective approach to make recognition more robust against common geometric variations and photometric noises. Representing a target object by a constellation of small image patches has an advantage of absorbing errors due to non-linear distortions in geometrical configuration of object parts across different same-class object instances. At the same time, however, it makes it more difficult to detect such small patches because their discriminative power tends to decrease by reducing the patch size. By semantically equivalent but visually dissimilar, we mean local object parts, sharing the same name but also being highly dissimilar in their visual appearance (e.g., a mouth in a face or a wing in an airplane). Such visual dissimilarity often renders the common detection approach with image-based similarity measures ineffective. For detecting such object parts, Epshtein and Ullman [4] recently proposed to exploit geometrically stable configural context. This approach characterizes the location of a visually dissimilar object (i.e., root fragment) with respect to its articulative configuration to a set of visually stable patches (i.e., context fragments) learned from the data. Experimentally, they demonstrate that correct detection of the visually dissimilar parts does improve recognition performance. The proposed method however does not allow any view-variations (i.e., scaling/rotation), preventing it from its practical applications.

Addressing this issue, this paper presents extensions of the Epshtein and Ullman's framework for robust detection of visually dissimilar objects with the presence of geometric domain variations up to similarity transformation. For this purpose, we integrate SIFT-based descriptor [9] for detecting context fragments using image-based similarity. A closed-form formula is also offered for generating scale- and rotation-invariant root candidates, exploiting geometric invariants of similar triangles without explicitly estimating an underlying domain transformation. An entropic feature selection is performed as a pre-process for improving overall performance by removing context fragments without significant image structures.

Another contribution of this work is to robustify the root location estimation from a collection of root candidates. The original formulation of a maximum likelihood estimator with i.i.d. Gaussians are highly sensitive to outliers caused by errors in finding context correspondences (see Figure 6 for example). To mitigate this, we propose an alternative maximum density framework that leads to a robust consensus voting. This framework models a root estimate as a statistical mode of a heteroscedastic kernel density estimator by interpreting each root candidate as an independent Gaussian sample. Resulting multi-modal density function is then analyzed with multiple seeds to detect the most significant mode by using variable bandwidth mean shift [3]. Ro-

bust estimation is possible by avoiding negative influences from outliers captured in non-significant modes.

The rest of this paper is organized as follows. Section 2 briefly introduces the Epshtein and Ullman's framework. Sections 3 and 4 describe our contributions in two successive steps of detecting context and root fragments, respectively. An overview of the overall algorithm is presented in Section 5. The proposed algorithms are experimentally evaluated for detecting various facial parts using the FERET database [10, 11]. Section 6 presents our experimental results which demonstrate the robustness of our approach. Finally, in Section 7 we conclude this paper by discussing our future work.

2. Related Work by Epshtein and Ullman

This section briefly introduces the semantically equivalent object detection framework proposed by Epshtein and Ullman [4], which serves us as a foundation of our work. The framework consists of two successive learning and detection phases. It takes a *root fragment* F and a set of training images T_r as an input. Root fragment F is a rectangular image patch which indicates a target object to be detected. The training image set T_r is then assumed to contain different instances of the same object.

The learning phase extracts a *context* C from T_r (context retrieval), where context C is defined as a set of N local image region called *context fragments*: $C_i, i = 1, .., N$. The main idea is to select a set of image patches that are geometrically consistent with the target root fragment as a context so that the root can be detected according to collective geometrical configuration on to such contexts. First step of this learning phase is to detect object parts visually similar to F. Normalized cross correlation (NCC) was used to exhaustively match F against T_r after applying difference of Gaussian (DoG) filtering. This procedure, called *DNCC*, results in a subset $\{I_k \in T_r, k = 1, .., K_i\}$. This subset contains K_i root-similar fragments F_k, centered at $\mathbf{x}_{fk}$, with their DNCC score exceeding a pre-defined threshold. In the next step, C is extracted from I_k as a collection of local image patches that are likely to co-occur with F_k in a stable geometric configuration. Local image patches with various sizes and locations $\mathbf{x}$ are exhaustively paired with F_k then posterior probability $P(F_k|Patch)$ and variance of coordinate differences $Var(\mathbf{x} - \mathbf{x}_{fk})$ are computed. After removing obviously unwanted patches using various thresholds, the patches are evaluated according to weights that are proportional to $P(F_k|Patch)$ and inversely proportional to $Var(\mathbf{x} - \mathbf{x}_{fk})$. Context C is then constructed by including patches that receive high weights, simultaneously maximizing the posterior and minimizing the configuration variance. For each C_i, the mean and variance of coordinate difference are also recorded.

The detection phase localizes root-similar fragments $F_n, n = 1, .., K$ in a set of test images T_e given the learned context C. Each context fragment C_i is first matched against T_e by using the same procedure as in the learning phase. Only images I_n that produce at least one confident match are considered for further processing. Finally, root fragment F_n for these K images are estimated with respect to the matched context fragments as a standard maximum likelihood estimator of F with a normal form of data likelihood $P(C_1, .., C_N|F)$, assuming independence of context fragments.

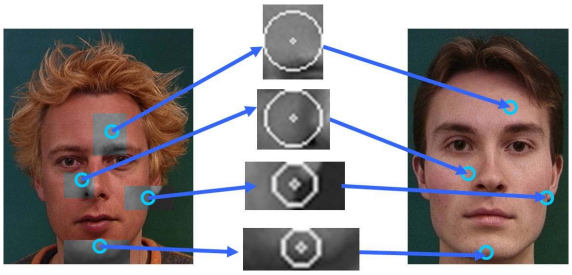

Figure 1. Process of detecting context fragments. Left: Selecting context fragments $C_i \in C_s$ using entropy. Center: Extracting SIFT-based descriptors $\mathcal{D}(C_i)$. Right: Matching C_i to $I_n \in T_e$, resulting in detected context fragments $\{C_{in}\}$.

In this paper, we generalize the detection phase of this framework toward more realistic image and geometric variations up to similarity transformation. For the learning phase, we use a reimplementation of the original algorithm as described above without any algorithmic modifications. For this study, we conducted the context learning of five facial parts F (i.e., mouth, nose, eye, ear, chin). We assemble T_r consisting of 20 images collected from the FAME database [15]. The above learning algorithm results in more than 100 context fragments C_i for each root fragment. Note that we consider the training set T_r and test set T_e to be disjoint sets while the original work consider them identical. This is also a straightforward extension of the original framework.

3. Detecting Context Fragments

3.1. Overview

This section summarizes our approach for localizing the context C in the test set T_e with the presence of image and geometric variations, given a pre-learned C for a root F. It is obvious to see that DNCC is suboptimal for our task and that more invariant image feature and/or similarity measures for matching are required. As was also suggested in [4], we adopt SIFT descriptors [9].

Figure 1 illustrates our approach. In order to improve the overall matching performance, we first perform a data-driven entropic feature selection, resulting in $C_s \subset C$ by

removing non-structural fragments detected by the original learning algorithm. For each selected context fragment $C_i \in C_s$ (*e.g.*, left nostril), we extract a SIFT-based descriptor $\mathcal{D}(C_i)$. Then $\mathcal{D}(C_i)$ is matched against SIFT descriptors computed at various locations of each test image in T_e. The location that yields the highest similarity (or lowest distance) is given as the final estimate, resulting in a set of detected context fragments $\{C_{in}, i = 1, .., N_n, n = 1, .., K\}$ for a subset of K test images $\{I_n\} \subset T_e$ each of which produces at least two confident matches. N_n denotes the number of context fragments detected for I_n.

We also evaluate three different similarity measures in order to improve the matching performance. Euclidean distance $ED(C_i, C_j)$, chi-square distance $\chi^2(C_i, C_j)$ [12], and earth mover's distance with L_1 ground distance $EMD\text{-}L_1(C_i, C_j)$ [8] are compared. Our experimental evaluation suggest that χ^2 performs better than other measures.

3.2. Data-Driven Context Selection with Entropy

Our pilot study indicated that the learning phase of the original framework can result in a large number of hard-to-detect context fragments which lack much spatial structure, resembling a homogeneous patch. This is due to the fact that the learning algorithm does not take into account such intensity statistics. These fragments therefore tend to result in low performance in detecting them on a test image.

Exploiting this observation, prior to the subsequent matching, we subject the context C to a data-driven feature selection using Shannon entropy [14] in order to remove such non-structural fragments. We first represent intensity distribution of C_i as a normalized histogram $P(C_i)$ with R bins. The entropy $H(P(C_i))$ is then defined,

$$H(P(C_i)) = -\sum_{r=1}^{R} P(C_i)_r \ln P(C_i)_r \tag{1}$$

We rank all C_i with $H(P(C_i))$. Then choose only C_is whose entropy exceeds a pre-defined threshold TH_H. This results in a subset $C_s \subset C$ of N_s selected context fragments such that $N_s \leq N$.

3.3. SIFT-based Descriptor

For each selected context fragment $C_i \in C_s$, we extract a SIFT-based descriptor $\mathcal{D}(C_i)$. The descriptor is computed following the Lowe's convention with 4×4 spatial sampling regions and 8 bins for the orientation histograms. We employ the public SIFT software developed at UCLA [1].

We compute $\mathcal{D}(C_i)$ at the center (x_{ci}, y_{ci}) of C_i unlike the original SIFT by Lowe [9], which extracts a set of descriptors at the selected interest-points (i.e., key-points) for each patch. This is because our pilot studies show that the success rate of locating C_i was too low when using the original SIFT for matching context fragments across different instances of the same object. Image variations beyond affine transformation, which occur naturally in such a situation, will prevent SIFT key-points from frequently corresponding to each other. Thus a matching with a single descriptor can outperform another matching with multiple descriptors extracted at such multiple key-points. However, in our descriptor design, the scale and rotation invariances become compromised because the descriptor is computed at a point other than the SIFT key-point. To address this issue, we compute a set of SIFT descriptors for B_s scales and B_r rotations for each C_i. Therefore, $\mathcal{D}(C_i)$ consists of a set of $B_s \times B_r$ SIFT descriptors. A similarity value of a pair of fragments C_i and C_j is then given by the highest value among those computed for all permutation of descriptor pairs.

3.4. Similarity Measurement of Local Descriptors

We compare three aforementioned similarity measures for matching the SIFT-based descriptors. Let $\mathcal{D}(C_i)_p$ be p_{th} bin of $\mathcal{D}(C_i)$ and similarly $\mathcal{D}(C_j)_p$ be p_{th} bin of $\mathcal{D}(C_j)$. The Euclidean distance between $\mathcal{D}(C_i)$ and $\mathcal{D}(C_j)$ from C_i and C_j, respectively, is then defined,

$$ED(C_i, C_j) = \|C_i - C_j\|_2 = \sqrt{\sum_{p=1}^{P} (\mathcal{D}(C_i)_p - \mathcal{D}(C_j)_p)^2} \tag{2}$$

We choose a match according to the lowest distance between $\mathcal{D}(C_i)$ and $\mathcal{D}(C_j)$.

Chi-square (χ^2) distance [12] compares a pair of descriptors $\mathcal{D}(C_i)$ with P number of bins. χ^2 is a hypothesis testing which evaluates null hypothesis that given $\mathcal{D}(C_i)$ and $\mathcal{D}(C_j)$ are dissimilar. Then, the χ^2 distance between $\mathcal{D}(C_i)$ and $\mathcal{D}(C_j)$ is defined,

$$\chi^2(C_i, C_j) = \sum_{p=1}^{p} \frac{(\mathcal{D}(C_i)_p - \mu_p)^2}{\mu_p}$$

$$\mu_p = \frac{\mathcal{D}(C_i)_p + \mathcal{D}(C_j)_p}{2} \tag{3}$$

We also choose a match according to the lowest distance between $\mathcal{D}(C_i)$ and $\mathcal{D}(C_j)$, assuming the corresponding hypothesis has the least possibility to be rejected.

EMD-L_1 [8] is an efficient algorithm to compute the earth mover's distance (EMD) [13] between multi-dimensional histograms. Unlike the typical bin-to-bin distances, such as ED and χ^2, EMD can absorb errors due to specific histogram binning as a cross-bin measure. By considering a specific case of L_1 ground distance, EMD-L_1 computes an EMD measure faster than the original formulation. In the following, we denote EMD-L_1 between $\mathcal{D}(C_i)$ and $\mathcal{D}(C_j)$ as EMD-$L_1(C_i, C_j)$.

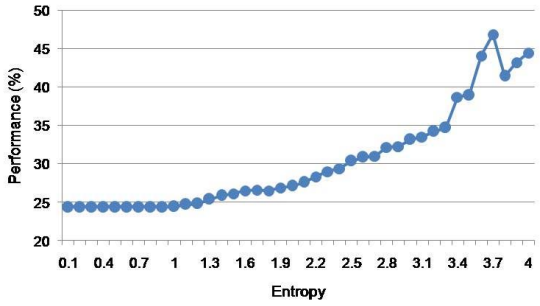

Figure 2. Data-driven entropic context selection. Detection rate of context fragment matching is shown as a function of the entropy threshold TH_H

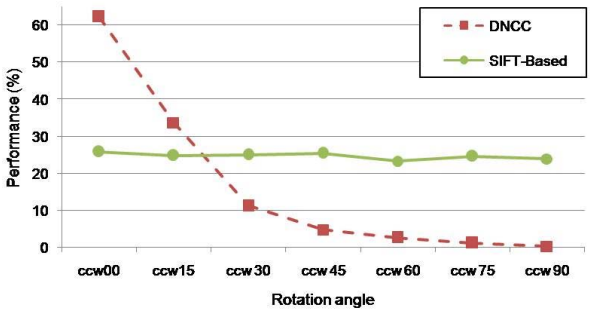

Figure 3. Comparing percentage of correctly locating C_i between SIFT-based matching and DNCC matching with rotation variance

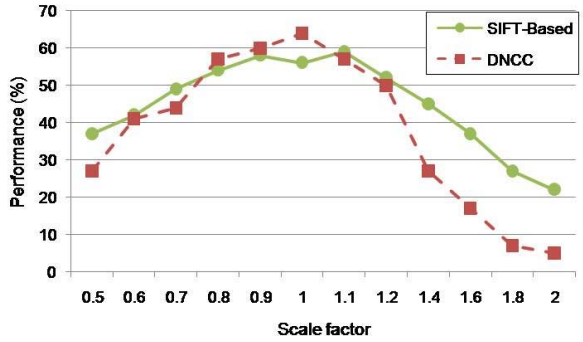

Figure 4. Comparing percentage of correctly locating C_i between SIFT-based matching and DNCC matching with scale variance

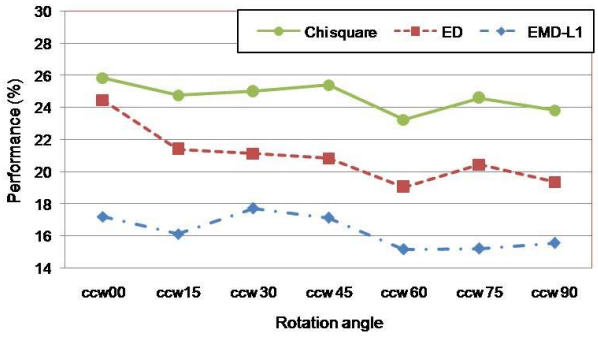

Figure 5. Comparing percentage of correctly locating C_i among different descriptor matching methods, χ^2, ED, and EMD-L_1.

3.5. Experimental Evaluations

This section presents our experimental validations of methods described in this section: 1) data-driven entropic context selection, 2) SIFT-based scale and rotation invariant context descriptors, and 3) similarity measures of the descriptors. These experiments use the contexts learned for detecting five facial parts F of mouth, nose, eye, ear, and chin as described in Section 2. In the following experiments, detection statistics are computed manually by visual inspection. We consider an estimated fragment to be a correct detection when a rectangle whose size is 130% of the estimated fragment fully contains the target region.

Figure 2 shows the results of our experiments for validating our data-driven context selection process. We conduct 1000 matching tests by locating 40 randomly chosen context fragments C_is for each of 5 roots in 5 randomly chosen test images $I_n \subset T_n \neq T_e$. The SIFT-based descriptor and χ^2 distance are used for matching. The figure displays correct detection rates as a function of the entropy threshold TH_H. The results show that increasing TH_H yields better detection rate with less C_is. For the following study, we set TH_H by 3.5, which results in $N_s \simeq 15$ where $N > 100$.

Figure 3 and 4 illustrates robustness of the SIFT-based descriptors in comparison to the DNCC matching used in [4]. For rotation variation, we conduct 17500 context matching tests for 7 rotations, 0°, 15°, 30°, 45°, 60°, 75°, and 90°, with 500 C_is for 5 Fs in 5 test images $\{I_n\}$. For scale variations, we have 1200 matching tests for 12 scales factors from 0.5 to 2.0. χ^2 distance was again used for matching. The results show that detection rates of the SIFT-based descriptors are more consistent than the DNCC matching across the scale and rotation variations.

Finally, figure 5 compares three different histogram similarity measures χ^2, ED, and EMD-$L1$ in the same rotation matching tests above. The results show that χ^2 consistently bests other two measures. Following this result, χ^2 will be used in the rest of our experiments.

4. Detecting a Root Fragment F

4.1. Overview

This section summarizes how we detect a root fragment F_n in a target test image I_n by using the detected context fragments C_{in} used as contextual clue. This part of our system consists of two successive steps: geometrically in-

variant candidate generation (GICG) and robust consensus voting using mean shift. Given a set of detected context fragments $\{C_{in}\}$, GICG generates a set of M scale- and rotation-invariant root candidates $F_{mn}, m = 1, ..., M$ in a test image $I_n \in T_e$. Taking F_{mn} as inputs, the following mean shift-based robust consensus voting yields a final localization of the root fragment F_n. The proposed algorithm parallels with RANSAC [6] however it offers more robust way for consensus voting without an explicit estimation of domain transformation. Details of these steps are described below.

4.2. Geometrically Invariant Candidate Generation

In [4], a root candidate F_{mn} is generated for each detected context fragment C_{in} in a test image I_n. Thus in this case the number of root candidates M is equivalent to the number of detected context fragments N_n. The center coordinate $\mathbf{x}_{fn}$ of each root candidate is estimated by,

$$\mathbf{x}_{fn} = \mathbf{x}_{cin} + \overline{\Delta \mathbf{x}_i} \tag{4}$$

where $\mathbf{x}_{cin}$ denotes the center coordinate of C_{in} and $\overline{\Delta \mathbf{x}_i}$ denotes the mean coordinate difference between the root and the i-th context fragment C_i learned in the training phase.

The above formula obviously fails when some domain variations are present as was also mentioned in [4]. This is because geometric constraints captured in the relative coordinate difference $\overline{\Delta \mathbf{x}_i}$ are not enough to uniquely determined the underlying transform (i.e., similarity transformation in $\mathbb{R}^2$ in our case). To address this, we estimate a root candidate from a *pair* of context fragments $(C_{in}, C_{jn}), i \neq j$, sampled from the set ofN_n detected fragments C_{in} in I_n. It is straightforward to see that two independent correspondences between learned and detected fragments can determine the similarity transform. When N_n is small, we can exhaust all 2-subsets of C_{in}, resulting in $M = \binom{N_n}{2}$ root candidates. When N_n is large, a random sampling of 2-subsets from C_{in} can be performed for $M < \binom{N_n}{2}$ times.

The following derives a closed-form formula for estimating $\mathbf{x}_{fn}$ from each pair (C_{in}, C_{jn}). Let $\mathbf{x}_f$, $\mathbf{x}_{ci}$, and $\mathbf{x}_{cj}$ denote the latent variables for center coordinates of the root fragment F and the corresponding context fragments C_i and C_j so that $\overline{\Delta \mathbf{x}_i} = \mathbf{x}_f - \mathbf{x}_{ci}$ and $\overline{\Delta \mathbf{x}_j} = \mathbf{x}_f - \mathbf{x}_{cj}$, respectively.

The main idea is to consider a pair of triangles $\mathcal{A}_1 = (\mathbf{x}_f, \mathbf{x}_{ci}, \mathbf{x}_{cj})$ and $\mathcal{A}_2 = (\mathbf{x}_{fn}, \mathbf{x}_{cin}, \mathbf{x}_{cjn})$ in $\mathbb{R}^2$. Then a specific assumption about the underlying domain variability can be interpreted as a corresponding geometrical relation between $\mathcal{A}_1$ and $\mathcal{A}_2$. This yields a set of equations representing geometric invariants that must hold true under the relationship. Finally, the set of equations can be solved to yield a closed-form formula for the unknown $\mathbf{x}_f$. Under our assumption that we allow up to the similarity transformation as the domain variability, therefore, triangles $\mathcal{A}_1$ and $\mathcal{A}_2$ are geometrically similar.

With different domain variability assumption, this general procedure can be readily extended by considering geometric invariants of a (S+1)-polygon pair. Such polygons can be constructed by sampling S-subsets from C_{in} where the value S is chosen such that S correspondences can sufficiently constrain the full degrees of freedom of the underlying transformation.

First, we re-describe the triangles with relative vectors from $\mathbf{x}_{ci}$ and $\mathbf{x}_{cin}$ so that $\mathcal{A}_1$ can be determined with information available during the detection phase,

$$\begin{aligned} \mathcal{A}_1 &= (\mathbf{c}_1, \mathbf{a}_1) = (\overline{\Delta \mathbf{x}_i},\ \overline{\Delta \mathbf{x}_i} - \overline{\Delta \mathbf{x}_j}) \\ \mathcal{A}_2 &= (\mathbf{c}_2, \mathbf{a}_2) = (\mathbf{x}_{fn} - \mathbf{x}_{cin},\ \mathbf{x}_{cjn} - \mathbf{x}_{cin}) \end{aligned}$$

Let $\mathbf{u}_1$ and $\mathbf{u}_2$ denote projections of $\mathbf{c}_1$ and $\mathbf{c}_2$ onto $\mathbf{a}_1$ and $\mathbf{a}_2$, respectively. Then we also re-describe $\mathbf{x}_f$ and $\mathbf{x}_{fn}$,

$$\begin{aligned} \mathbf{x}_f &= \mathbf{x}_{ci} + \mathbf{c}_1 = \mathbf{x}_{ci} + \mathbf{u}_1 + \mathbf{v}_1 \\ \mathbf{x}_{fn} &= \mathbf{x}_{cin} + \mathbf{c}_2 = \mathbf{x}_{cin} + \mathbf{u}_2 + \mathbf{v}_2 \end{aligned} \tag{5}$$

where

$$\begin{aligned} \mathbf{u}_1 &= k_1 \mathbf{a}_1 \\ k_1 &= \mathbf{c}_1 \cdot \mathbf{a}_1 / \|\mathbf{a}_1\|^2 \\ \mathbf{v}_1 &= \mathbf{c}_1 - \mathbf{u}_1 \\ \mathbf{u}_2 &= k_2 \mathbf{a}_2 \end{aligned} \tag{6}$$

Furthermore, $\mathbf{v}_2$ can be written as a function of normal vectors $\mathbf{n}_1$ and $\mathbf{n}_2$ to $\mathbf{a}_1$ and $\mathbf{a}_2$,

$$\mathbf{v}_2 = \text{sign}(\mathbf{c}_1 \cdot \mathbf{n}_1) \|\mathbf{v}_2\| \mathbf{n}_2 \tag{7}$$

where $\mathbf{a}_1 = (a_{11}, a_{12})$, $\mathbf{a}_2 = (a_{21}, a_{22})$, $\mathbf{n}_1 = (a_{12}, -a_{11})/\|\mathbf{a}_1\|$, and $\mathbf{n}_2 = (a_{22}, -a_{21})/\|\mathbf{a}_2\|$. Because $\mathcal{A}_1$ and $\mathcal{A}_2$ are similar triangles, the projection factor and ratio of vector lengths are invariant, yielding

$$k_2 = k_1 \tag{8}$$

$$\|\mathbf{v}_2\| / \|\mathbf{u}_2\| = \|\mathbf{v}_1\| / \|\mathbf{u}_1\| \tag{9}$$

Plugging in (8) and (9) to (5), (6), and (7) results in the final result that extends (4),

$$\mathbf{x}_{fn} = \mathbf{x}_{cin} + k_1 \mathbf{a}_2 + \text{sign}(\mathbf{c}_1 \cdot \mathbf{n}_1) \|\mathbf{c}_1 - \mathbf{u}_1\| \frac{\|\mathbf{a}_2\|}{\|\mathbf{a}_1\|} \mathbf{n}_2 \tag{10}$$

4.3. Robust Consensus Voting using Mean Shift

In [4], the center location of root fragment F_n is estimated from M root candidates F_{mn} using maximum likelihood estimation (MLE). They model a probability distribution of F_n being found at coordinate $\mathbf{x}$ given an observed context fragment C_{in} as a 2D Gaussian centered at

Figure 6. Illustration of difference between maximum likelihood and maximum density approaches. Left: maximum likelihood estimation. Right: maximum density-based robust consensus voting using mean shift. White dots indicate center locations of root candidates F_{mn} for mouth case.

$\mathbf{x}_{fn}$ with sample covariance of coordinate difference $\Delta\mathbf{x}_i$ between F and C. Under the independence assumption, the closed-form global maximizer $\hat{\mathbf{x}}$ of the data likelihood $P(C_{1n}, .., C_{in}, .., C_{N_n n}|F_n)$ is analytically derived.

This approach is sensitive to errors in detecting context fragments. Figure 6 demonstrates this shortcoming for the mouth detection case. Such detection errors results in a significant number of largely false root candidates while majority of candidates are still estimated correctly. The MLE solution cannot yield correct root estimate in this case.

To address this issue, we propose an alternative probabilistic model using a density estimator framework. We interpret each root candidate F_{mn} as an independent sample with an associated uncertainty in the Gaussian form. The heteroscedastic kernel density estimator $f(\mathbf{x})$ is then defined by summing these Gaussians,

$$f(\mathbf{x}) = \sum_{m=1}^{M} \mathcal{N}(\mathbf{x}; \mathbf{x}_{fmn}, \Sigma_m) \tag{11}$$

where $\mathbf{x}_{fmn}$ is the estimated center coordinate for F_{mn} and Σ_m is the corresponding covariance matrix, derived from sample covariances with i-th and j-th context fragments used to estimate $\mathbf{x}_{fmn}$, $\Sigma_m = \frac{\text{Cov}(\Delta\mathbf{x}_i)+\text{Cov}(\Delta\mathbf{y}_i)}{2}$.

We define a *mode* of this density function to be an estimated root center $\mathbf{x}_{fn}$. Such a mode indicates the *maximum density location*. It is intuitively a location with strongest evidence with many other estimates with similar values. This model uses the exactly same amount of information in comparison to [4] but results in a multi-modal distribution (i.e., Gaussian sum) unlike the uni-modal one (i.e., Gaussian product) in [4]. Robust estimation is possible by choosing a mode with correct estimates thus avoiding the negative influence from outlier candidates. Figure 6 illustrates the advantage of this approach in comparison to the MLE solution shown in the left image of the figure. As a local maximum density estimator, we use variable-bandwidth mean shift proposed by Comaniciu [3]. This iterative mode-seeking algorithm is provably convergent to a density mode in the vicinity of the initialization $\mathbf{x}_{init}$ similar to the gradient-descent but without the need for tuning the nuisance learning rate parameter. The iterator is defined,

$$\mathbf{m}_v(\mathbf{x}) = H(\mathbf{x}) \sum_{m=1}^{M} w_m(x)\Sigma_m^{-1}\mathbf{x}_{fmn} - \mathbf{x} \tag{12}$$

where $H(\mathbf{x})$ denotes the data-weighted harmonic mean of the bandwidth matrices at $\mathbf{x}$ such that

$$H^{-1}(\mathbf{x}) = \sum_{m=1}^{M} w_m(x)\Sigma_m^{-1}$$

The weight $w_m(x)$ represents the influence from m-th Gaussian component at $\mathbf{x}$ normalized over all the components

$$w_m(x) = \frac{|\Sigma_m|^{-\frac{1}{2}} \exp(-\frac{1}{2}(\mathbf{x}-\mathbf{x}_{fmn})^T\Sigma_m^{-1}(\mathbf{x}-\mathbf{x}_{fmn}))}{\sum_{m=1}^{M}|\Sigma_m|^{-\frac{1}{2}} \exp(-\frac{1}{2}(\mathbf{x}-\mathbf{x}_{fmn})^T\Sigma_m^{-1}(\mathbf{x}-\mathbf{x}_{fmn}))}$$

.

In order to robustly detect the most significant mode among others, we use the following voting scheme with multiple initializations. First, we set M initializations by $\mathbf{x}_{init,m} \leftarrow \mathbf{x}_{fmn}$. Second, we perform M mean shift procedures, resulting in M convergences $\mathbf{y}_m$. Third, we perform voting by grouping $\{\mathbf{y}_m\}$ into D clusters then choosing the cluster $\mathcal{C}$ with the most members. Finally, we set the final estimate $\mathbf{x}_{fn}$ by the average: $\mathbf{x}_{fn} \leftarrow \frac{1}{N_d}\sum_{d\in\mathcal{C}}\mathbf{y}_d$

5. Algorithm Overview

INPUT: A context $C = \{C_i\}$ trained for a root fragment F, average root location $\Delta\mathbf{x}_i$ and its covariance Cov($\Delta\mathbf{x}_i$), and a test image set T_e

OUTPUT: A set of estimated center location $\mathbf{x}_{fn}$ in F_n for a test image $I_n \in T_e$

ALGORITHM

1. Add C_i into C_s if $H(P(C_i)) > TH_H$ [Section3.2]
2. Extract $\mathcal{D}(C_i), C_i \in C_s$ with B_r rotations and B_s scales [Section3.3]
3. Detect C_i in each image $I_n \in T_e$ by finding a location with minimum $\chi^2(C_i, C_j)$ (3), resulting in detected context fragments $\{C_{in}\}$
4. Generate M root candidates $\{F_{mn}\}$ in I_n for every context fragment pairs (C_i, C_j) from $\{C_{in}\}$ by using (10) [Section4.2]
5. Estimate the location of root fragment $\mathbf{x}_{fn}$ from $\{F_{mn}\}$ by using (12) [Section4.3]

6. Experiments

This section presents the results of testing feasibility and robustness of detecting F in an image with geometric variation using various combination of the mentioned component algorithms: robust consensus voting with mean shift (MS), geometrically invariant candidate generation (GICG), the original Epshtein and Ullman's system with the MLE formulation (E+U), and DNCC matching (DNCC) [4] Similar to our previous experiments, we consider detecting five facial subparts, mouth, right eye, nose, ear, and chin, as root fragments F. Prior to our detection experiments, corresponding context C for each root F is also learned using the method described in Section 2.

6.1. Data

For our detection experiments, we use 1000 faces which are collected from FERET database [10, 11]. These faces exhibit some facial expressions (i.e., choosing from FERET database's "fa" and "fb" subsets equally). After the collection, each image is cropped so that parts other than face (*e.g.* shoulder) is removed. The cropped image is then resized to 150×200 (width $\times$ height) using bi-cubic interpolation and then is rotated counter clockwise (CCW) $15°$, $30°$, $45°$, $60°$, and $75°$ for the rotation invariance test and also resized with 0.5 and 1.5 scale for the scale invariance test. The resulting set of images constitute our testing image set T_e.

6.2. Methods

The detection phase is divided into three steps where different methods can be applied. The first step of the detection phase is to locate C_i using SIFT-based or DNCC matching. The second step is to generate candidates $F_{mn}, m = 1, ..., M$ in a test image $I_n \in T_e$ by GICG. The third step is then to estimate the root location of F_n in I_n using MS or MLE. As was described in Section 3.5, we consider F_n correctly estimated when the region around F_n with the size of 130% as the size of F has all the contents of F.

6.3. Results

First we perform an experiment to investigate effectiveness of GICG and MS (GICG + MS) while using the original DNCC matching on data without scaling and rotation variations. We conducted the root detection matching tests only with non-rotated data set $T_{CCW00} \subset T_e$. As you can see in Figure 7, (GICG + MS) bests (E+U) with at most 28% difference using the same parameter settings for the context fragments C_i, the number of C_i used, and threshold for DNCC matching. The difference in performance is due to the ability of (GICG + MS) to effectively filter out many outliers due to errors in finding corresponding context fragments.

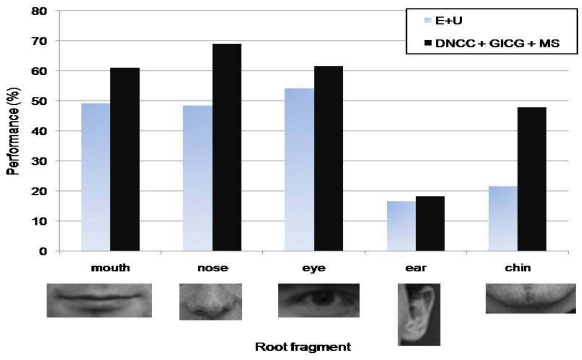

Figure 7. Comparing percentage of correctly locating C_i between DNCC+GICG+MS and E+U with no-rotation data set.

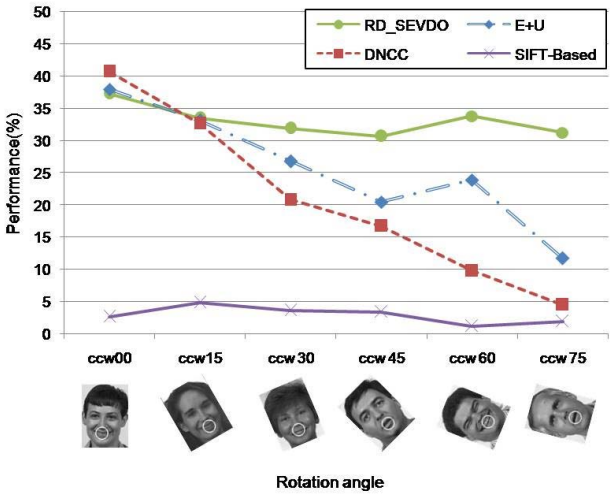

Figure 8. Comparing percentage of correctly locating C_i among SIFT-based + GICG + MS (RD_SEVDO), E+U, DNCC matching and SIFT-based matching for different rotation variance.

For the invariance test of rotation and scale, Figure 8 and Figure 9 show the results of 5000 tests each for rotation and scale with 1000 test images for each F. Comparing to the baseline SIFT-based and DNCC matching as well as (E+U), our system, (SIFT-based + GICG + MS) labeled "RD_SEVDO", yields consistent detection performance over scale and rotation variations. Note that the baseline SIFT-based and DNCC matching are performed as direct template matching with these image-based similarity measures for detecting the root fragments without using the contexts.

7. Discussion and Future Work

This paper presents our robust detection framework for locating semantically equivalent but visually dissimilar ob-

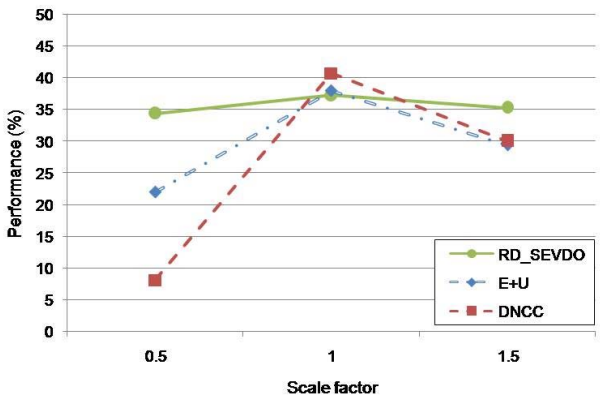

Figure 9. Comparing percentage of correctly locating C_i among SIFT-based + GICG + MS (RD_SEVDO), E+U, and DNCC matching for different scale variance.

ject parts with the presence of scale and rotation view-variations. Our experiments demonstrate our approach's significant improvements over the original Epshtein and Ullman's system in terms of detection performance and of robustness against the domain scaling and rotation. This allows us to consider practical usage of the overall framework toward more realistic application scenarios, contributing to improve the general part-based object recognition paradigm. As our future work, we plan to experimentally evaluate the proposed algorithm with articulated objects other than faces in both 2D and 3D domains. The mean shift-based maximum density voting is not restricted to the domain's dimensionality. Furthermore, the proposed GIGC can be readily extended to 3D domains. Another interesting future work is to extend the learning phase of the Epshtein and Ullman's framework toward scale and rotation invariance so that context fragments can be learned from images with arbitrary view-variations. The SIFT-type invariant features can also be incorporated into the learning process, facilitating to select better context locations that contain certain image structures that current learning phase ignores.

References

[1] http://vision.ucla.edu/ vedaldi/code/sift/sift.html.

[2] E. Bart, E. Byvatov, and S. Ullman. View-invariant recognition using corresponding object fragments. In *Proc. of European Conf. Computer Vision*, 2004.

[3] D. Comaniciu. An algorithm for data-driven bandwidth selection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(2):281–288, 2003.

[4] B. Epshtein and S. Ullman. Identifying semantically equivalent object fragments. In *Proc. of IEEE Conf. Computer Vision and Pattern Recog.*, 2005.

[5] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. of IEEE Conf. Computer Vision and Pattern Recog.*, 2003.

[6] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. of the ACM*, 24:381–395, 1981.

[7] B. Heisele, T. Serre, M. Pontil, T. Vetter, and T. Poggio. Categorization by learning and combining object parts. In *Neural Information Processing Systems*, 2001.

[8] H. Ling and K. Okada. An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29:840–853, 2007.

[9] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Computer Vision*, 60(2):91–110, 2004.

[10] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[11] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss. The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5):295–306 1998.

[12] J. Puzicha, T. Hofmann, and J. Buhmann. Non-parametric similarity measures for unsupervised texture segmentation and image retrieval. In *Proc. of IEEE Conf. Computer Vision and Pattern Recog.*, 1997.

[13] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *Int. J. Computer Vision*, 40:99–121, 2000.

[14] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, july, october 1948.

[15] M. B. Stegmann, B. K. Ersboll, and R. Larsen. Fame – a flexible appearance modeling environment. *Medical Imaging*, 22(10):1319–1331, 2003.