Detecting and Localizing 3D Object Classes using Viewpoint Invariant Reference Frames

Matthew Toews and Tal Arbel
Centre for Intelligent Machines
McGill University, Montreal, Canada
{mtoews,arbel}@cim.mcgill.ca

Abstract

In this paper, we investigate detection and localization of general 3D object classes by relating local scale-invariant features to a viewpoint invariant reference frame. This can generally be achieved by either a multi-view representation, where features and reference frame are modeled as a collection of distinct views, or by a viewpoint invariant representation, where features and reference frame are modeled independently of viewpoint. We compare multi-view and viewpoint invariant representations trained and tested on the same data, where the viewpoint invariant approach results in fewer false positive detections and higher average precision. We present a new, iterative learning algorithm to determine an optimal viewpoint invariant reference frame from training images in a data-driven manner. The learned optimal reference frame is centrally located with respect to the 3D object class and to image features in a given view, thereby minimizing reference frame localization error as predicted by theory and maintaining a consistent geometrical interpretation with respect to the underlying object class. Modeling and detection based on the optimal reference frame improves detection performance for both multiview and viewpoint invariant representations. Experimentation is performed on the class of 3D faces, using the public color FERET database for training, the CMU profile database for testing and SIFT image features.

1. Introduction

Vision research has long focused on how to understand, learn and recognize the three dimensional world from two dimensional projections. A primary question that arises is that of representation: what is a suitable representation for modeling, detecting and localizing 3D object classes in 2D imagery? Recent computer vision literature addressing general object class detection and localization has advocated models based on local image features [12, 25]. At the heart of these models lies the notion of a geometrical reference frame defined in the image plane, that is related to the underlying 3D object class in a geometrically meaningful way. The reference frame serves as a mechanism for grouping image features arising from the same underlying object class instance, and thereby defining detection. Examples include bounding boxes [10], feature constellations [12], individual features [11, 8], etc.

The majority of approaches to modeling object class appearance to date have focused on single viewpoints, and as a result reference frames used are often single viewpoint in nature, i.e. they do not remain geometrically consistent with the underlying 3D object class with a change in viewpoint. Bounding boxes change in size and shape with in-plane rotation, 2D feature configurations collapse and individual features disappear with in-depth rotation. Such reference frames do not lend themselves easily to learning and detecting 3D object classes from arbitrary viewpoints. For a reference frame to remain geometrically consistent over change in viewpoint, it must represent a property of the underlying 3D object class that is invariant to perspective projection arising from viewpoint change, i.e. a viewpoint invariant [4]. Examples include 3D object centroids which project to 2D points [18], 3D volumetric primitives such as geons [2] or spheres which project to 2D circles, 3D line segments which project to 2D lines [22, 4].

In this paper, we consider modeling of 3D object classes from arbitrary viewpoints based on viewpoint invariant reference frames. There are two major representations by which this can be accomplished [3], the multi-view representation and the viewpoint invariant representation. The multi-view representation links image features to reference frames in a set of distinct views around the object class of interest, thereby explicitly modeling the variable of viewpoint. The viewpoint invariant representation links image features directly to a viewpoint invariant reference frame, thereby effectively marginalizing the variable of viewpoint.

Our contribution is to address two important, open questions relating to modeling 3D object classes from viewpoint

978-1-4244-1631-8/07/$25.00 ©2007 IEEE

invariant reference frames. First, is it preferable to model in terms of a set of distinct views (i.e. the multi-view representation) or in terms of a single model over the entire viewpoint range (i.e. the viewpoint invariant representation)? Second, can an optimal invariant reference frame be derived for an object class in an iterative, data-driven manner? We address these questions as follows. First, we provide a side-by-side comparison of multi-view and viewpoint invariant representations in the context of supervised 3D face modeling and detection, in terms of detection performance. We hypothesize that the linking features directly to the 3D geometry of the object class via a viewpoint invariant representation instead of to specific views will improve detection. Second, we propose an iterative algorithm to learn an optimal viewpoint invariant reference frame, by iteratively 1) learning a viewpoint invariant model from training images with labeled reference frames, then 2) re-estimating reference frame labels from the learned model. We hypothesize that this learning process will converge to a reference frame that both 1) remains geometrically consistent with the underlying 3D object class and 2) minimizes the error in predicting the reference frame geometry from image features.

The remainder of this paper is organized as follows. In Section 2 we review related work on object class detection, particularly recent local feature-based approaches. In Section 3, we review the theory behind the specific approach we adopt in learning multi-view and viewpoint invariant representations. In Section 4, we present experimentation in the context of 3D face learning and detection. Learning is based on the standard color FERET face database [1], which contains a large number of different subjects with viewpoint labels, and detection is based on the public CMU profile database [7] containing faces in arbitrary viewpoints amid clutter. Results show that the viewpoint invariant representation is superior in terms of the average precision measure [10], as the multi-view representation is more prone to false positives. Iterative learning results in an optimal reference frame that remains geometrically consistent with 3D faces, and detection performance is improved using the optimal reference frame for both multi-view and viewpoint invariant representations. In Section 5 we conclude with a discussion.

2. Related Work

Our work follows the trend in recent computer vision literature to model object class appearance in terms of local scale-invariant features [15, 16, 5, 13]. Although efficient and effective detectors have been constructed based on features such as Haar wavelets [26], scale-invariant features are interesting due to their high degree of invariance to nuisances, including partial occlusion, illumination changes and in-plane geometrical deformations such as translation, scale and orientation changes. Additionally, invariant features can be efficiently detected and matched without requiring an explicit search over parameters of geometrical transforms under which they are invariant. While invariant features can often be directly matched between different images of the same textured object, i.e. object recognition/detection [14], they cannot be generally matched directly between different instances the same abstract object class, e.g. different cars or faces.

Probabilistic modeling and machine learning have been used to address the task of general object class detection for a variety of different classes, e.g. vehicles, animals, etc [12, 9, 20, 8, 6]. Such models describe the appearance of an object class in terms of a set of local features, including their occurrence statistics, appearances and geometries (i.e. image location, orientation and scale) with respect to a geometrical reference frame. As reference frames used are often specific to fixed viewpoints, e.g. 2D feature constellations [12], they are generally ineffective for detecting and localizing general 3D object classes from arbitrary viewpoints [27], where inter-feature geometrical relationships and individual feature appearances change significantly with changes in viewpoint.

Representations used to model general 3D object class appearance over viewpoint change can be generally described as multi-view or viewpoint invariant in nature [3]. The multi-view representation maintains a sampling of distinct single viewpoint models around the object of interest, and detection is accomplished by fitting a new image to the nearest modeled view. The viewpoint invariant representation models features with respect to a viewpoint invariant reference frame, and detection is accomplished inferring the reference frame from image features in a manner independent of viewpoint. In the computer vision literature, both multi-view [21] and viewpoint-invariant [22] representations have been proposed to address 3D object class detection and localization from arbitrary viewpoints [1]. Although the multi-view/viewpoint-invariant distinction is helpful for understanding representations, elements of both representations can be combined. For example, a viewpoint invariant reference frame can be used in both 1) a viewpoint invariant representation or 2) as the basis for modeling individual views in a multi-view representation.

3. Modeling 3D Object Classes using Viewpoint Invariant Reference Frames

Our work builds on the viewpoint invariant modeling approach of Toews and Arbel [22], which we outline in Section 3.1. In Section 3.2 we describe the implications of using viewpoint invariant reference frames with multi-view

[1] Note that geometry-free "bag-of-feature" models do provide a mechanism for learning appearance from arbitrary viewpoints but cannot be used directly to localize or enumerate object class instances within images [9, 20].

and viewpoint invariant representations. In Section 3.3, we propose an iterative algorithm for learning an optimal invariant reference frame in a data-driven manner.

3.1. Viewpoint Invariant Reference Frames

Modeling appearance over viewpoint change can be accomplished by linking features to a viewpoint invariant reference frame, which maintains a consistent a consistent geometrical interpretation with respect to the underlying 3D object class from arbitrary viewpoints. In the context of 3D object classes, such an invariant can be referred to as an OCI (object class invariant) [22], a viewpoint invariant reference frame that is uniquely defined for each instance of a 3D object class. Figure 1 illustrates an example of an OCI in the form of a 3D line segment for the modeling the class of 3D face images from scale-invariant features. The 3D line segment is uniquely defined for each 3D face from the base of the nose to the forehead, a definition which is independent of viewpoint. Such a reference frame is suitable for representing the appearance of object classes such as faces which are typically viewed from a coronal plane (i.e. rarely from over/underhead views), and thus bear a distinct orientation component.

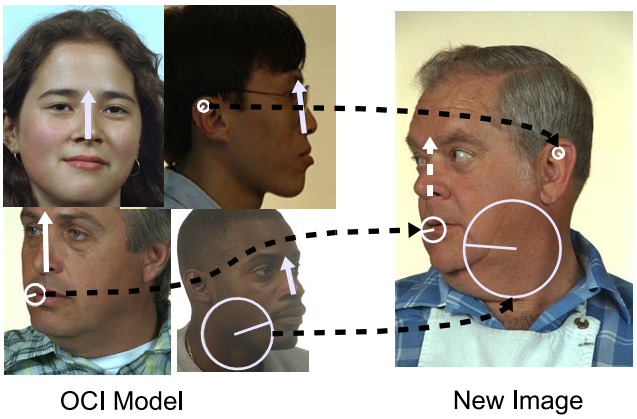


Figure 1. Illustrating modeling the object class of 3D faces in terms of scale-invariant features (white circles) and a viewpoint invariant reference frame (solid white arrows). The reference frame, defined here a line segment from the base of the nose to the forehead, maintains a consistent geometrical interpretation with respect to the face from arbitrary viewpoints. It can thus be used to group scale-invariant image features in images taken from arbitrary viewpoints, for the purpose of face detection and localization. A probabilistic model (left) can be learned from labeled reference frames in training images taken at arbitrary viewpoints. Model instances can then be detected and localized in a new image (right) taken at an arbitrary viewpoint, based on detected model features (dashed black lines) that agree on the reference frame (dashed white arrow). Note that reference frame shown here 1) exploits the symmetry of faces allowing mirror feature correspondence and 2) is not designed for overhead/underhead views.

3D object class appearance can be described in terms of an OCI o and a set of scale-invariant image features $\{m_i\}$. A feature is denoted as $m_i : \{m_i^g, m_i^a, m_i^b\}$, and consists of variables of geometry m_i^g, appearance m_i^a and occurrence m_i^b. Feature geometry $m_i^g : \{\sigma_i, \theta_i, x_i\}$ is a scale-invariant geometrical description of the feature in an image in terms of its scale σ_i, orientation θ_i and (row,col) image location x_i. Feature appearance m_i^a represents the image content within the region specified by the image geometry and can be represented in a number of ways, e.g. principle components [24] or orientation histograms [15]. Feature occurrence m_i^b represents the presence or absence of a feature. The OCI is denoted as $o : \{o^g, o^b\}$ consisting of variables of geometry o^g and occurrence o^b. Geometry o^g is a viewpoint invariant reference frame, which in the case of a line segment is equivalent to a scale-invariant model feature geometry. Note that this OCI definition is similar to that of a model feature but lacks an appearance component, as an OCI is not directly observable from image data.

The relationship between OCI o and model features $\{m_i\}$ can be described probabilistically as:

$$p(o|\{m_i\}) = \frac{p(o)p(\{m_i\}|o)}{p(\{m_i\})} = p(o)\frac{\prod_i p(m_i|o)}{p(\{m_i\})}, \quad (1)$$

where the first equality results from Bayes rule and the second from the assumption of conditional feature independence given the OCI. With the conditional independence assumption, modeling focuses on term $p(m_i|o)$ defining the relationship between an individual feature and the OCI as:

$$p(m_i|o) = p(m_i^a|m_i^b)p(m_i^b|o^b)p(m_i^g|o^b, o^g), \quad (2)$$

under the assumptions of conditional independence of features appearances/occurences $\{m_i^a, m_i^b\}$ and geometries $\{m_i^g\}$ given the OCI o, and independence of feature appearance m_i^a and OCI occurrence o^b given feature occurrence m_i^b. Term $p(m_i^a|m_i^b)$ represents feature appearance given presence, and can be modeled as Gaussian. Term $p(m_i^b|o^b)$ represents the binomial probability of feature occurrence given reference frame occurrence. Term $p(m_i^g|o^b, o^g)$ represents the residual error in predicting the reference frame geometry from the feature geometry, and can be modeled as Gaussian. Note that the scale parameters are treated in the log domain, and location parameters are normalized by reference frame scale.

Learning requires estimating the parameters of the terms in equation (2). This can be done using a supervised learning technique in order learn a model from natural imagery from arbitrary viewpoints. First, OCIs are manually labeled in a set of training images by drawing a line segment on the image, for example from the base of the nose to the forehead as in Figure 1, and scale-invariant features are automatically extracted in all training images. With labeled

OCI and extracted features, learning proceeds by identifying clusters of features that agree in terms of their appearances and their geometries with respect to the OCI, where each such cluster represents a single underlying model feature m_i. To identify clusters, each extracted feature m_i is treated as a potential model feature. A feature m_j is said to agree geometrically with m_i if, when normalized according to their geometries m_i^g and m_j^g, their respective OCIs o_i^g and o_j^g differ by less than a scale-invariant threshold T^g in scale, orientation and location. Features m_j that agree geometrically are considered as events $o^{b=1}$, and those that do not agree are considered as events $o^{b=0}$. Note that T^g represents the maximum acceptable error in predicting the OCI geometry, and thus a single empirically determined threshold is applicable for all features. Two features are said to agree in terms of appearance if the difference between their appearances m_i^a and m_j^a is less than an appearance threshold T_i^a. Features m_j that agree in appearance are considered as events $m_i^{b=1}$ and those that do not agree are considered as events $m_i^{b=0}$. Unlike the global geometrical threshold T^g, the appearance threshold T_i^a is feature-specific and determined by image content of individual features. It is set such that the likelihood ratio $\frac{p(m_i^{b=1}|o^{b=1})}{p(m_i^{b=1}|o^{b=0})}$ of geometrically agreeing vs. disagreeing features is maximized. Note that this ratio can be considered a measure of feature distinctiveness [9], and appearance thresholds are thus set to maximize the distinctiveness of model features.

Detecting and localizing object class instances in a new image can be done as follows. Features are first extracted in the new image and matched to model features. An image feature m is said to match a model feature m_i if the difference in their appearance representations is less than the learned appearance threshold T_i^a. Each model-to-image match implies the geometry of an OCI o^g in the new image, and clusters of similar geometries o^g suggest the presence of a valid OCI. Different hypotheses o_i^g and o_j^g are considered as belonging to the same cluster if their difference is less than the same global geometrical threshold T^g used in model learning. The hypotheses that a particular OCI cluster results from a true OCI instance $o^{b=1}$ or noise $o^{b=0}$ can be tested using a Bayes decision ratio [12]:

$$\begin{aligned}\gamma(o^g) &= \frac{p(o^g, o^{b=1}|\{m_i\})}{p(o^g, o^{b=0}|\{m_i\})}, \\ &= \frac{p(o^g, o^{b=1})}{p(o^g, o^{b=0})} \prod_i \frac{p(m_i|o^g, o^{b=1})}{p(m_i|o^g, o^{b=0})}. \end{aligned} \quad (3)$$

In equation (3), term $\frac{p(o^g, o^{b=1})}{p(o^g, o^{b=0})}$ is a constant representing the prior ratio of valid vs. invalid OCI o^g occurrences, and $\frac{p(m_i|o^g, o^{b=1})}{p(m_i|o^g, o^{b=0})}$ represents the likelihood ratio of a true vs. false feature match.

3.2. Multi-view or Viewpoint Invariant Modeling?

Given that features are to be modeled relative to a viewpoint invariant reference frame, it is possible to adopt either the multi-view or viewpoint invariant representation using the OCI modeling technique in the previous section. Which should be used? The multi-view approach is arguably more prevalent in the literature [21, 18], possibly because of simplicity: individual view models do not require invariant reference frames and can generally make use of a wide variety of single viewpoint techniques. The choice of reference frame aside, however, the difficult with the multi-view representation is as follows: local features arising from images of 3D object class typically persist over a range of viewpoint [15], the particular range depending on the specific image feature. When the range of a particular feature overlaps between adjacent views of a multi-view model, which is often the case, the same image feature is linked to distinctly different reference frames in adjacent views. When a multi-view model is used for detection, the same image features will thus produce strong, distinctly different hypotheses as to the geometry of the reference frame in the image. Indeed, a major focus of multi-view modeling is developing strategies to cope with false detections or imprecise localization resulting from the same image features supporting different hypotheses as to the object class in the image [21]. Additional, special learning algorithms have been proposed to identifying modeling features which can be shared across views [23].

The viewpoint invariant model avoids this difficulty as follows: image features are learned over the viewpoint range in which they best predict the reference frame geometry on a feature-by-feature basis, and not according to a fixed view sampling. For this reason, we hypothesize that a viewpoint invariant representation will generally result in fewer false positive detections, and thus improved detection performance.

3.3. Deriving Optimal Invariant Reference Frames

What is an optimal viewpoint invariant reference frame? Can one be derived in an iterative, data-driven manner? Viewpoint invariant reference frames used are often specified arbitrarily, e.g. as an object centroid [18] or as a line segment along the nose in face images [22]. These definitions have intuitive 3D interpretations and lend themselves to supervised learning from natural imagery, as labels can be specified in arbitrary viewpoints (one could even guess at their locations from rear views of a head). But are such reference frames optimal for detection? One could argue to the contrary, for the reason that error in predicting the reference frame geometry image features increases with distance between the two in the image, as illustrated in 2. An optimal reference frame would thus minimize the expected distance

between observed image features and the reference frame, while the nasally defined OCI for instance is clearly non-central for oblique and profile views of a face.

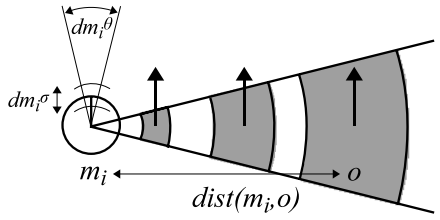


Figure 2. The error in predicting the location of the reference frame o from an image feature m_i increases with the distance $dist(m_i, o)$ between the two. The three grey arc regions illustrate the localization error for three different values of $dist(m_i, o)$, which is a function of errors in feature scale dm_i^σ and orientation dm_i^θ and $dist(m_i, o)$.

Ideally, fully unsupervised learning could derive an optimal invariant reference frame from data, however unsupervised learning is a challenging task even in the single viewpoint case. Approaches to learning models over viewpoint changes typically employ a degree of external supervision. The multi-view representation, for example, requires viewpoint information and is learned from a viewsphere sampled at regular angular intervals, for each of set of different object class instances [21]. Images must be sorted and aligned according to viewpoint, and a suitable sampling of viewpoints must be determined, making it difficult to learn from natural imagery taken from arbitrary viewpoints. A viewpoint invariant representation does not require explicit viewpoint knowledge and can be learned from natural images taken from arbitrary viewpoints using manual labeled OCIs [22].

Here, we extend learning in order to determine an optimal viewpoint invariant in an iterative, data-driven fashion: starting from an initial, coarse OCI labeling as in the previous section, we propose alternately learning features $\{m_i\}$ from reference frames labels $\{o_j\}$ as in equation (2), then re-estimating OCI labels $\{o_j\}$ from learned features $\{m_i\}$ by maximizing the decision ratio in (3). The initial labeling step constitutes a degree of supervision that could be removed via a partially supervised learning approach where a small portion of data is labeled, then propagated by determining initial coarse feature correspondences between different faces and views. We hypothesize that this iterative learning process will converge to a stable OCI definition that is geometrically consistent with the underlying 3D object class and result in improved detection performance.

4. Experimentation

The goals of experimentation are twofold. First, we compare the viewpoint invariant and multi-view representations in terms of object class detection and localization performance. Second, we show how the iterative learning algorithm can be used to identify optimal viewpoint invariant reference frames. The details of the experimental setup are as follows.

Image features: Although a variety of different scale-invariant features can be used, we use the scale-invariant feature transform (SIFT) technique [15] based on an implementation provided by the author. SIFT features have been shown to perform well in comparison with other approaches in terms of detection repeatability [19] and appearance distinctiveness [17].

Training data: Model learning is based on the publicly available color FERET face image database [1], consisting of images of 994 unique subjects of various ethnicity, age, gender, taken from various viewpoints, illumination conditions, with/without glasses, etc. As FERET images are labeled according to viewpoint, they provide a good basis for multi-view modeling and therefore comparison of the multi-view and viewpoint invariant representations. For the purpose of training, we randomly select 497 different subjects (half of all FERET subjects), and for each subject, we randomly select a viewpoint from the range of -90 to 90 degrees (i.e. left profile to right profile) provide by the FERET database. This results in a total of 497 images, which are processed in grey scale at a resolution of 256x384 pixels. Each image results 150-300 SIFT features. Training OCI geometries are manually labeled from the base of the nose to the forehead.

Testing data: Detection performance is evaluated on a subset of images from the CMU profile database [7], containing a total of 95 faces. The database is a challenging detection test set containing face images under arbitrary viewpoints amid a high degree of background clutter. We select a subset of testing faces which are of higher resolution, as low resolution faces (i.e. < 40 pixels) produce few SIFT features and are thus difficult to detect. Ground truth OCI locations were labeled as line segments from the base of the nose to the forehead, as in training images.

4.1. Multi-view vs. Viewpoint Invariance

Given that a viewpoint invariant reference frame exists for modeling 3D object classes, is it better to maintain a set of distinct views (multi-view representation) or ignore viewpoint altogether (viewpoint invariant representation)? Here we compare multi-view and viewpoint invariant representations learned from precisely the same training labels and image features, in terms of detection performance. For the purpose of multi-view learning, we quantize the range of

viewpoint into 3 distinct views: frontal, oblique and profile, as illustrated in Figure 3. The multi-view representation is generated by learning three distinct models from the images in each of the three view ranges, and the viewpoint invariant representation is generated by learning a model from all 497 images, both via the learning process described in Section 3.1.

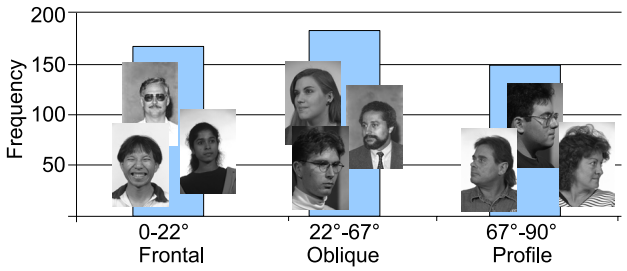


Figure 3. A histogram of the three viewpoint ranges used in multi-view modeling: frontal, oblique and profile. The 497 test images are approximately equally distributed across all views. Note that as the model exploits the mirror symmetry of the face, a 180 degree range of viewpoint can be covered by the three views.

Detection was performed by fitting the respective models to features extracted in the test images, as described in Section 3.1. To suppress multiple detection hypotheses arising from the same face, all hypotheses in a proximity of a hypothesis with a locally maximal Bayes decision ratio (3) were removed, where proximity was defined by geometrical OCI agreement threshold T^g as in the discussion on model learning. Note that the detection output was considered for all three view models of the multi-view representation, while the viewpoint invariant model produces only a single detection output. Figure 4 illustrates the precision-recall curves of viewpoint invariant and multi-view. As hypothesized, the viewpoint invariant representation outperforms the multi-view representation in terms of the average precision (AP) measure, defined by the arithmetic mean of the precision evaluated at recall increments of 0.1 [10].

The reason the viewpoint invariant representation achieves superior detection performance because of the higher false positive rate of the multi-view approach. This is particularly true when the viewpoint of a testing image falls in between the two training views of the multi-view representation, as illustrated in Figure 5. In this situation, the same image features fit well to different view models, producing strong, distinctly different geometrical interpretations as to the geometry of the object class instance. Although this ambiguity can be resolved to a degree in the multi-view context [21], cases arise in natural imagery where is is generally difficult to establish wether neighboring hypotheses arise from two distinct object class instances or simply different interpretations from different views, as

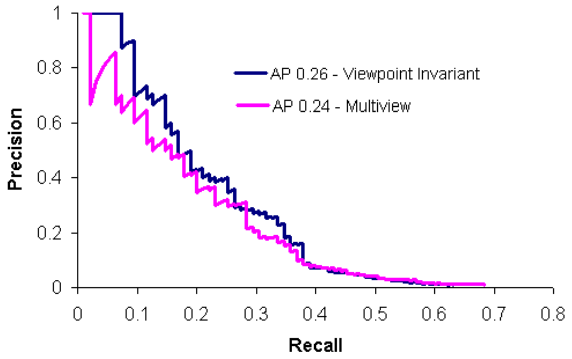


Figure 4. Precision-recall curves for viewpoint invariant and multi-view detection. Viewpoint invariant detection generally outperforms multi-view detection, with respective average precisions of 0.26 and 0.24.

illustrated in Figure 6. The difficulty is generally less prevalent with the viewpoint invariant representation, where each feature effectively serves as its own best view.

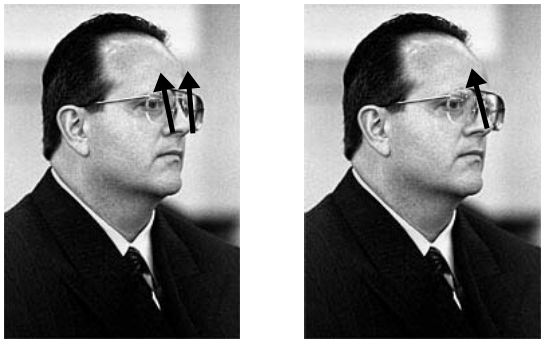


Figure 5. The viewpoint of the face shown falls somewhere between the oblique and profile views of the multi-view model. As a result, the oblique and profile views of the multi-view representation produce two strong, distinct interpretations as to the geometry of the object class instance, as shown in A). The viewpoint invariant representation produces a single strong interpretation, as shown in B), as image features are related directly to the geometry of the object class instance and not to fixed views. Note that the frontal view of the multi-view representation did not produce a significant detection hypothesis for this image.

4.2. Determining an Optimal Viewpoint Invariant

Can an optimal viewpoint invariant reference frame be derived in a data-driven manner? Starting with the FERET face images initially labeled as in the previous section, we perform the iterative process described in Section 3.3. OCIs appear to converge in the training images after approximately 30 iterations. For frontal face training images, the

Figure 6. The two overlapping faces to the right illustrate a situation where it is difficult to determine wether multiple detection hypotheses are the result of different interpretations of the same object class instance or two distinct object class instances.

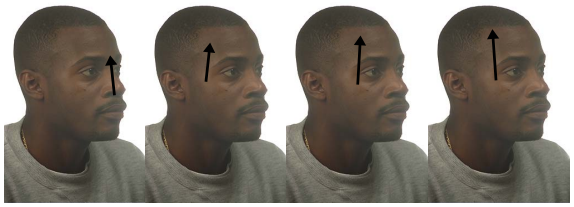
re-learned OCI labels become slightly larger in scale but do not change significantly in orientation or location. In all oblique and profile images, however, the OCI labels retreat noticeably from the nose back to the cheeks, as illustrated in Figure 7. These new labels, determined by an iterative, data-driven process, are consistent with the projection of a single 3D line segment central to the human head in the image plane. As such, they minimize the distance between image features and the projected OCI location over a 180 degree range of viewpoint.

In order to evaluate the new optimal OCI definition, we re-run detection trials. The difficulty, however, is that the ground truth OCI geometries in the testing images must be relabeled according to the new OCI definition. As the new optimal OCI definition has been automatically determined, the human labeler him/herself must first learn it from training image examples before producing ground truth labelings in the testing images. The detection results are shown in Figure 8, where the optimal OCI results in improved detection performance over the initial manual OCI in terms of average precision for both multi-view and viewpoint invariant representations.

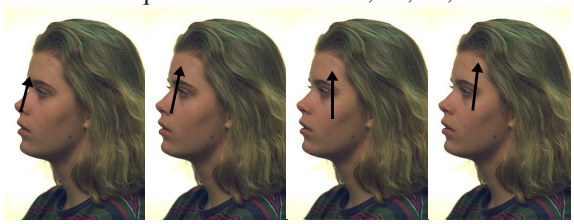
5. Discussion

In this paper, we investigate open questions regarding the viewpoint invariant representation of 3D object class appearance from scale-invariant image features, particularly relating to the use of a viewpoint invariant reference frame consistent with the geometry of the 3D object class. When modeling using viewpoint invariant reference frames, is it preferable to use a multi-view or a viewpoint invariant representation? A comparison of detection performance shows that the viewpoint invariant representation outperforms the multi-view representation in terms of average precision, as

Frontal view: iterations 0, 10, 20, 30



Oblique view: iterations 0, 10, 20, 30



Profile view: iterations 0, 10, 20, 30

Figure 7. Illustrating the result of iteratively learning and re-estimating OCI labels in training images, for 0, 10, 20 and 30 iterations. In iteration 0, all OCIs are manually initialized as line segments from the base of the nose to the forehead. Little change occurs for OCIs after 30 iterations in frontal views, which are already approximately central to image features arising from the face. In quarter and profile views, OCI locations recede to the cheeks, minimizing the average distance to image features characteristics of these views (e.g. ears, cheeks, eyes). Note that the OCIs in all views remain consistent with 3D geometry of the object class, corresponding to the 2D projections of the same 3D line segment located within the head.

the multi-view representation is prone to a higher number of strong, false detection solutions. Can an optimal viewpoint invariant reference frame be derived in an data-driven manner? An iterative algorithm is proposed which converges to a stable reference frame central to the underlying 3D object class and image features, thereby minimizing the error in reference frame localization as predicted by theory. Both multi-view and viewpoint invariant models trained using the optimally derived reference frame show improvement in detection performance.

Open questions remain. Is it important to model 3D object class appearance using viewpoint invariant reference frames? As mentioned, a variety of single viewpoint models that could potentially be used to model views in a multi-view representation. It is reasonable to expect, however, that the primary difficulty with the multi-view model would

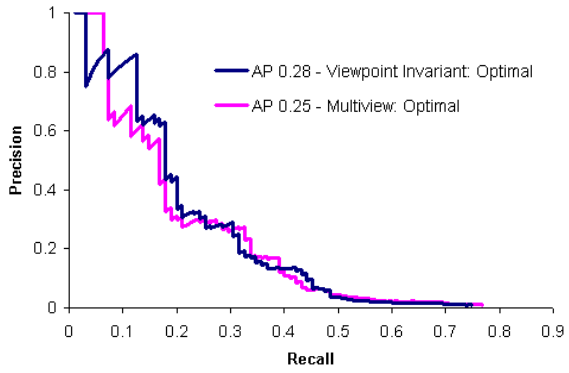


Figure 8. Precision-recall curves using the optimal OCI for both viewpoint invariant and multi-view detection. Using the optimal OCI definition improves detection for both viewpoint invariant and multi-view representations from the initial definition, which show average precision (AP) improvements of $0.26 \rightarrow 0.28$ for viewpoint invariant detection and $0.24 \rightarrow 0.25$ for multi-view detection.

still be manifested, i.e. the same image feature being related to multiple, different views. Do all object classes generally exhibit meaningful viewpoint invariant reference frames? It would seem so for rigid object classes such as cars or 3D scenes, a collection of reference frames would likely be required to efficiently encode deformable or articulated object classes. Can viewpoint invariant reference frames serve as a basis for fully unsupervised learning of 3D object class appearance? Our work shows that iterative model learning and reference frame re-estimation converges to a stable result in training images from arbitrary viewpoints. The key is in automatically establishing several initial reference frame labels sufficiently similar to one another to drive learning to convergence.

References

[1] Feret face database. www.itl.nist.gov/iad/humanid/colorferet. 2, 5

[2] I. Beiderman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94(2):115–147, 1987. 1

[3] I. Beiderman and P. C. Gerhardstein. Recognizing depth-rotated objects: Evidence and conditions for 3d viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance*, 19:1162–1182, 1993. 1, 2

[4] J. Burns, R. Weiss, and E. Riseman. The non-existence of general-case view-invariants. In *Geometric Invariance in Computer Vision*, pages 120–131, 1992. 1

[5] G. Carneiro and A. Jepson. Multi-scale phase-based local features. In *CVPR*, volume 1, pages 736–743, 2003. 2

[6] G. Carneiro and D. G. Lowe. Sparse flexible models of local features. In *ECCV*, volume III, pages 29–43, 2006. 2

[7] CMU Face Group. Frontal and profile face databases. http://vasc.ri.cmu.edu/idb/html/face/. 2, 5

[8] D. Crandal, P. Felzenswalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *CVPR*, volume 1, pages 10–17, 2005. 1, 2

[9] G. Dorko and C. Schmid. Selection of scale-invariant parts for object class recognition. In *ICCV*, pages 634–640, 2003. 2, 4

[10] M. Everingham, A. Zisserman, C. Williams, and L. Van Gool. The pascal visual object classes challenge 2006 (voc2006) results. Technical report, September 2006. 1, 2, 6

[11] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *ICCV*, 2005. 1

[12] R. Fergus, P. Perona, and A. Zisserman. Weakly supervised scale-invariant learning of models for visual recognition. *IJCV*, 71(3):273–303, 2006. 1, 2, 4

[13] T. Kadir and M. Brady. Saliency, scale and image description. *IJCV*, 45(2):83–105, November 2001. 2

[14] D. Lowe. Local feature view clustering for 3d object recognition. In *CVPR*, pages 682–688, 2001. 2

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, November 2004. 2, 3, 4, 5

[16] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 60(1):63–86, October 2004. 2

[17] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–, October 2005. 5

[18] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *ECCV*, pages 69–81, 2004. 1, 4

[19] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *IJCV*, 37(2):151–172, June 2000. 5

[20] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their localization in images. In *ICCV*, pages 370–377, 2005. 2

[21] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. Van Gool. Towards multi-view object class detection. In *CVPR*, 2006. 2, 4, 5

[22] M. Toews and T. Arbel. Detection over viewpoint via the object class invariant. In *ICPR*, volume 1, pages 765–768, 2006. 1, 2, 3, 4, 5

[23] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *CVPR*, pages 762–769, 2004. 4

[24] M. Turk and A. P. Pentland. Eigenfaces for recognition. *CogNeuro*, 3(1):71–96, 1991. 3

[25] S. Ullman, M. Vidal-Naquet, and E. Sali. Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, (5):682–687, 2002. 1

[26] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, volume 1, pages 511–518, 2001. 2

[27] M. Weber, W. Einhauser, M. Welling, and P. Perona. Viewpoint-invariant learning and detection of human heads. In *Int. Conf. on Automatic Face and Gesture Recognition*, pages 7–20, 2000. 2