Detection Over Viewpoint via the Object Class Invariant

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

Abstract

In this article, we present a new model of object class appearance over viewpoint, based on learning a relationship between scale-invariant image features (e.g. SIFT) and a geometric structure that we refer to as an OCI (object class invariant). The OCI is a perspective invariant defined across instances of an object class, and thereby serves as a common reference frame relating features over viewpoint change and object class. A single probabilistic OCI model can be learned to capture the rich multimodal nature of object class appearance in the presence of viewpoint change, providing an efficient alternative to the popular approach of training a battery of detectors at separate viewpoints and/or poses. Experimentation demonstrates that an OCI model of faces can be learned from a small number of natural, cluttered images, and used to detect faces exhibiting a large degree of appearance variation due to viewpoint change and intra-class variability (i.e. (sun)glasses, ethnicity, expression, etc.).

1. Introduction

In this article, we consider object class appearance modeling for the tasks of detection and localization. Although models of specific object appearance can often be learned from a single image [13, 15, 9, 14], object class appearance modeling remains an active area of research [16, 1, 7, 6, 5], due to the difficulty in learning the wide range of appearance variability characterizing an object class.

We focus in particular on the parts-based appearance representation [16, 1, 2, 7, 6], which has recently enjoyed a wave of interest due the maturity of local feature detectors. Generic scale-invariant features [10, 11, 8], for instance, can be efficiently and robustly extracted from images of a wide variety of objects, in the presence of illumination changes and in-plane geometrical deformations such as scale, rotation and translation. Appearance models based on such features can be efficiently learned and used to detect object instances in the presence of occlusion. A significant drawback of many such models, however, is that they attempt to infer or detect a stable 2D configuration of features, and are thus inherently single-viewpoint in nature. Extending such models to account for viewpoint change requires a battery of detectors at different viewpoints/poses [16, 12], adding complexity to model size, learning and fitting (detection) [16].

Our contribution is a new model of object class appearance, designed specifically to address the difficulty of representing appearance over viewpoint change, which we refer to as the OCI (object class invariant) model. We define an OCI as a perspective invariant common to all members of an object class. As such, an OCI serves as a common reference frame relating scale-invariant features over viewpoint. Although such an invariant is difficult to extract directly [3], we show that a probabilistic model relating scale-invariant features to an OCI can be learned from natural, cluttered imagery, and used to infer an object instance in the form of an OCI in a new image, all in the presence of viewpoint change, see Figure 1.

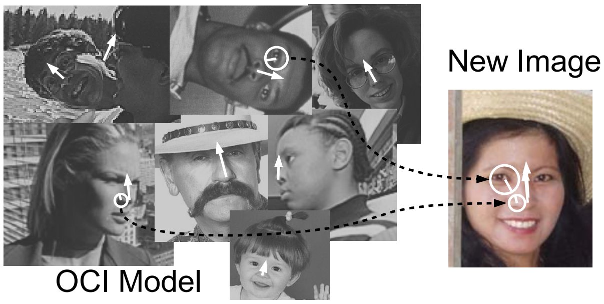

Figure 1. An OCI (shown here as a white arrow) is a perspective invariant defined across all instances of an object class, i.e. faces in the CMU database [4]. A probabilistic model relating scale-invariant features (shown as white circles) to an OCI can be used to infer an OCI in a new image.

The remainder of this article is organized as follows: we describe the OCI model in Section 2, followed by experimentation consisting of viewpoint-invariant detection in

Section 3 and a discussion in Section 4. Although the OCI model can generally be applied to a variety of object classes, we carry through the example of face modeling for the purpose of illustration.

2. The Object Class Invariant Model

In this section, we derive the object class invariant (OCI) model, designed specifically to capture the rich multimodal nature of object class appearance in the presence of viewpoint change. We present the components and the probabilistic formulation of the OCI model, and describe how such a model can be learned and used to detect OCI instances from natural, cluttered imagery.

2.1. Components of the OCI Model

The OCI model is based on generic scale-invariant features [10, 11, 8], which can be automatically extracted from images and consist of two distinct quantities: 1) a 4-parameter geometrical structure $g : \{x, y, \sigma, \theta\}$ describing the x,y location, scale and rotation of a feature within an image and 2) a vector a describing the image intensity appearance at g. In addition, we consider a binary variable b representing the presence (or absence) of a feature. A model feature is thus denoted as $m : \{m^b, m^g, m^a\}$, representing the presence, location and appearance of a scale-invariant feature related to the object class.

Within the context of this paper, we adopt an OCI in the form of a line segment, a perspective invariant that shares the same geometrical representation as the scale-invariant feature. This has the implication that a single scale-invariant feature is sufficient to infer an OCI location. The OCI is thus denoted as $o : \{o^b, o^g\}$ representing the presence and location of an OCI within an image. Note that o does not contain an appearance component, as it is not directly observable from an image, but rather inferred via m.

2.2. Probabilistic OCI Formulation

The OCI model is designed quantify the probability of o given a set of N model features $\{m_i\}$. Assuming that $\{m_i\}$ are conditionally independent given o, this probability can be expressed using Bayes rule as:

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

where $p(o)$ is a prior over OCI location and occurrence, and $p(m_i|o)$ is the likelihood of feature m_i given o. Our model focuses principally on the likelihood term $p(m_i|o)$, which can be expressed as:

$$\begin{aligned} p(m_i|o) &= p(m_i^a, m_i^b|o)p(m_i^g|o) \\ &= p(m_i^a|m_i^b)p(m_i^b|o^b)p(m_i^g|o^b, o^g), \quad (2) \end{aligned}$$

under the assumptions that m^a and m^b are statistically independent of m^g given o, and that m^a and o are statistically independent given m^b.

Appearance likelihood $p(m_i^a|m_i^b)$ is represented as a multivariate Gaussian distribution in an appearance space and parameterized by mean and covariance μ_i^a, Σ_i^a. $p(m_i^b|o^b)$ is the probability of model feature occurrence given hypothesis occurrence, represented as a discrete multinomial distribution with event count parameters $\pi_i = \{\pi_i^1, \ldots, \pi_i^4\}$. Geometry likelihood $p(m_i^g|o^b, o^g)$ models the residual error of a 4-parameter linear transform from feature to OCI geometry $m_i^g \rightarrow o^g$, and is represented as a Gaussian distribution with mean and covariance parameters μ_i^g, Σ_i^g. In order to characterize geometrical error in a scale-invariant manner, scale is transformed logarithmically, and translation is normalized by OCI scale.

2.3. Learning, Detection and Localization

Learning in a natural scenario requires efficiently identifying a small set of features $\{m_i\}$ common to an object class, while rejecting the large majority which arise from unrelated clutter, possibly with the help of weak supervision. These features must be both informative regarding o and non-redundant, i.e. do not repeat information in other features. This way, the number of model features N is kept from growing indefinitely and the independence assumption in equation (1) remains valid.

To replicate such a scenario, we distance ourselves from training images that are aligned or sorted by viewpoint. Instead, we consider a set of natural images, containing feature observations $\{m_i^a, m_i^g\}$ and a manually labeled hypothesis o^g. Labeling o^g represents a form of weak supervision, and is easily done by tracing a line segment corresponding to the OCI on the image. Unsupervised learning could be performed in the absence of such labeling, at the cost of increased search time, particularly for the difficult imagery we use - we leave this for future work. Scale-invariant features are extracted and represented using the SIFT technique [10], based on an efficient implementation provided by the author, although a variety of other techniques could be used. Briefly, SIFT features are extracted as maxima/minima in a difference-of-Gaussian scale space pyramid, determining feature geometry m_i^g. The SIFT appearance representation m_i^a is a 128-value vector, corresponding to bins of a histogram of image first derivatives quantized into 8x4x4=128 bins over orientation and (x,y) position.

Learning involves estimating model parameters based on a set of data vectors of the form $\{m_i^a, m_i^g, o^g\}$ with missing data $\{m_i^b, o^b\}$, and proceeds via the following two-step process. Step (1): a set of samples $\{o^b\}$ is generated from spatial likelihood $p(m_i^g|o^b, o^g)$, after initializing parameters μ_i^g, Σ_i^g to reasonable values. Intuitively, this step involves evaluating the spatial agreement of feature pairs $\{m_i, m_j\}$ wrt OCI geometry o^g. Step (2): a set of samples $\{m_i^b\}$ is generated from appearance likelihood $p(m_i^a|m_i^b)$. Intuitively, this step involves evaluating the appearance agreement of feature pairs $\{m_i, m_j\}$, and is done by setting parameters μ_i^a, Σ_i^a to maximize likelihood ratio $\frac{p(m_i^{b=1}|o^{b=1})}{p(m_i^{b=1}|o^{b=0})}$. This process could iterate by estimating μ_i^g, Σ_i^g and restarting from (1), we found only a single iteration was necessary.

Once model parameters have been learned, features with low $\frac{p(m_i^{b=1}|o^{b=1})}{p(m_i^{b=1}|o^{b=0})}$ can be discarded, as they provide little information regarding o. In order to reduce feature redundancy, we discard m_i that have high mutual information with other model features $\{m_j\}$ given o, in a manner similar to [2]. Mutual information estimation is noisy, however, as $p(m_i^{b=1}|o^b)$ is typically poorly sampled due to the rare occurrence of $m_i^{b=1}$. As a result, significant dependencies still exist between database features, artificially reducing the likelihood in equation (2). We adopt a heuristic technique based on the assumption that spatial overlap of features m_i with respect to o implies dependence, to more accurately estimate the joint probability.

For the purpose of detection and localization, we are interested in evaluating whether a set of data observations in a new image are the result of a true OCI or random noise, i.e. $o = \{o^g, o^{b=1}\}$ or $\bar{o} = \{o^g, o^{b=0}\}$. These hypotheses can be compared via a Bayes decision ratio:

$$\gamma(o) = \frac{p(o|\{m_i\})}{p(\bar{o}|\{m_i\})} = \frac{p(o)}{p(\bar{o})} \prod_{i=1}^{N} \frac{p(m_i|o)}{p(m_i|\bar{o})}, \qquad (3)$$

where $\gamma(o) > 1$ indicates the presence of an OCI and ratio $\frac{p(o)}{p(\bar{o})}$ is a context-specific factor which determines the false detection rate. Optimization requires determining $o^g = \underset{o^g}{\text{argmax}} \{\gamma(o)\}$, which involves a search over all combinations of one-to-one pairings from model features to data observations, or to no observation at all (in which case $\frac{p(m_i|o)}{p(m_i|\bar{o})} \approx 1$). Although generally intractable, this search can be constrained by identifying clusters of features producing similar hypotheses o^g.

3. Experimentation

The goal of experimentation was to demonstrate the feasibility of 1) learning an OCI model from a small set of natural, cluttered training images, and 2) using the model to detect new class instances, all in images exhibiting a wide range of appearance variability due to viewpoint change and intra-class variation. For the purpose of experimentation, we chose to model the class of face images due to the abundance of raw data, although the OCI model is generally applicable to any object classes containing detectable scale invariant features, such as cars, etc. Training and testing data consisted of 180 examples of faces of different people taken from a variety of viewpoints, from the internet and the CMU profile database [4], in addition to a set of 43 negative image of scenes not containing faces. The positive examples contained a high degree clutter, and exhibited a wide variety of appearance variability due to viewpoint change, (sun)glasses, expressions, race, etc., see Figure 2.

OCIs were defined and labeled as a line segments from the nose tip to the forehead, as in Figure 1, extreme side or rear views of the face can be labeled by guessing the approximate projection of the OCI, when the landmarks are not visible. Such an OCI is sufficient for modeling faces over a 360 degree range of viewpoint in the axial plane, assuming an orthographic projection model (i.e. object size is small relative to the distance between object and camera), although the OCI magnitude vanishes in overhead views. Modeling object appearance over an entire viewsphere could be accomplished by considering several different OCIs corresponding to perpendicular 3D line segments in the world - two such OCIs would limit the maximum magnitude discrepancy of a single OCI to $\frac{1}{\sqrt{2}}$, for example - we leave this for future work.

Model learning involved $\approx 16,000$ features, of which $\approx 12,000$ were deemed to be uninformative wrt o and discarded, in addition to $\approx 1,300$ which were deemed to be redundant, resulting in a model of $\approx 2,800$ features. This demonstrates learning in the presence of significant clutter. Each detected OCI consisted of a unique combination of $10-20$ model features from different training images, highlighting the ability of the OCI model to represent a large range of appearance modes. Model learning is quick, on the order of minutes, and detection on the order of seconds for images of size $\approx 300 \times 200$ pixels.

Detection trials were performed in a leave-one-out manner - a model was trained using the entire training set except for one image, which was then used to test the model, for each image in the training set. A detected hypothesis was considered successful if it fell within a scale-dependent threshold of the labeled OCI, i.e. a difference in scale, orientation and (x,y) translation of log(1.5) octaves, 20 degrees, and $0.5/\sigma$ pixels, where σ is the scale of the labeled OCI. In addition, a mechanism of non-maximal suppression was applied to remove potential hypotheses in a neighborhood around hypothesis maxima. Figure 2 illustrates the result of detection trials. The ROC curve was based on a total of 180 valid detections and 26,475 false positives, and rises quickly to a maximum detection rate of 81% - this rate is conservative, as several near-solutions were labeled as false, i.e. that of Pete Sampras kissing the trophy.

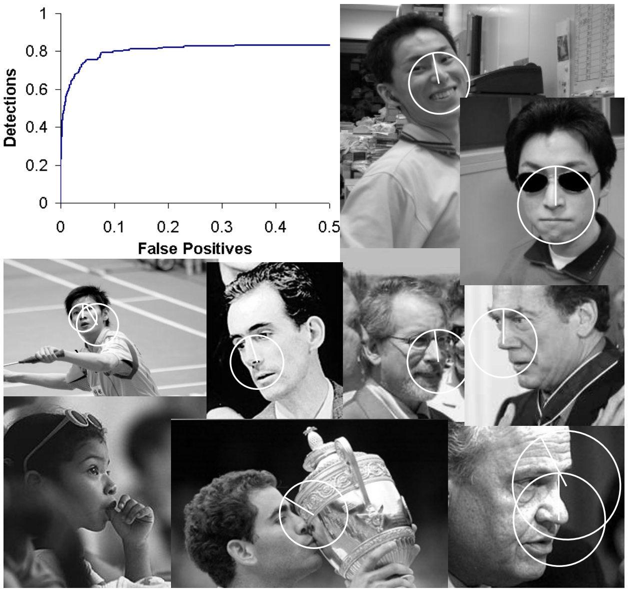

Figure 2. Detection results on a database of 180 face images, including images from the CMU profile database [4], containing a wide range of viewpoint and appearance variability. The ROC curve in the upper left corner plots the detection vs. false positive characteristic of the OCI model as $\frac{p(o)}{p(\bar{o})}$ is varied. The white circles overlaying the images show OCI detections, false positives, and missed detections for a particular value of $\frac{p(o)}{p(\bar{o})}$.

4. Discussion

In this article, we presented a new model of object class appearance over viewpoint, based on a perspective invariant defined across instances of an object class which we refer to as an OCI. A single OCI model is capable of representing viewpoint change, in-plane translation, rotation and scale, in addition to intra-class appearance variability, providing an alternative to the battery of single viewpoint models required by many approaches to represent viewpoint change.

In comparing models of object class appearance, it is important to consider the context surrounding learning and detection. Models suitable for learning on thousands of aligned images are arguably superior in terms of detection performance [12], but the degree of supervision required is daunting. Other models are interesting in that they learn with very little supervision [7], but typically require simplifying assumptions, such as a small fixed number of model features (i.e. 3-7) with unimodal appearance (i.e. no sunglasses), and single viewpoint data (i.e. car rear). The OCI model falls somewhere in between - the degree of supervision is low, as images are unaltered and contain significant clutter, yet a single OCI model is able to learn and detect the rich multimodal appearance of a difficult face image set, in the presence of viewpoint change.

The current implementation showed the feasibility of OCI modeling from a viewplane where the OCI magnitude remains approximately constant wrt object size. Future work will involve testing multiple OCIs for detection over complete viewspheres. We intend to apply the OCI framework to different object classes, using different types of invariant features. Finally, we are currently pursuing OCI model learning in the unsupervised context.

References

[1] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *PAMI*, 26(11):1475–1490, 2004.

[2] E. Bart, E. Byvatov, and S. Ullman. View-invariant recognition using corresponding object fragments. In *ECCV04*, pages 152–165, 2004.

[3] J. Burns, R. Weiss, and E. Riseman. View variation of point-set and line-segment features. *PAMI*, 15(1):51–68, 1993.

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

[5] G. Dorko and C. Schmid. Selection of scale-invariant parts for object class recognition. In *ICCV03*, pages 634–640, 2003.

[6] L. Fei-Fei, R. Fergus, and P. Perona. A bayesian approach to unsupervised one-shot learning of object categories. In *ICCV03*, pages 1134–1141, 2003.

[7] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR03*, pages 264–271, 2003.

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

[9] D. Lowe. Local feature view clustering for 3d object recognition. In *CVPR01*, pages 682–688, 2001.

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

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

[12] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *ECCV04*, pages 69–81, 2004.

[13] P. Moreels, M. Maire, and P. Perona. Recognition by probabilistic hypothesis construction. In *ECCV04*, pages 55–68, 2004.

[14] A. Pope and D. G. Lowe. Probabilistic models of appearance for object recognition. *IJCV*, 40(2):149–167, 2000.

[15] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3d object modeling and recognition using affine-invariant patches and multi-view spatial constraints. In *CVPR03*, page 272 277, 2003.

[16] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *CVPR04*, pages 762–769, 2004.