Affine Stable Characteristic based Sample Expansion for Object Detection

Ke Gao [1], Yongdong Zhang [1], Wei Zhang [1, 2], Shouxun Lin [1]

[1]Institute of Computing Technology, Chinese Academy of Sciences,
Beijing, China, 100190

[2]Graduate University of the Chinese Academy of Sciences, Beijing, China, 100080

{kegao, zhyd, zhangwei, sxlin}@ict.ac.cn

ABSTRACT

Generating better object model from automatic expanded samples is an effective approach to improve the performance of object detection. However, most existing methods either don't work well with limited relevance images in corpus, or result in redundant features and the decrease of detection speed. In this paper, we propose a novel method called Affine Stable Characteristic to generate an object feature model using only one object sample. By integrating affine simulation with stable characteristic mining, a compact and informative object model is generated with high robustness to viewpoint and scale transformations. For characteristic mining, two new notions, *Global Stability* and *Local Stability*, are introduced to calculate the robustness of each object feature from complementary hierarchies. And they are combined to generate the final object feature model. Experiments show that our novel method is capable of detecting objects in various geometric and photometric transformations, while only acquiring one sample image. In a compiled dataset composed of many famous test sets, the detection accuracy can be improved 35.8% compared with traditional methods at rapid on-line speed. The proposed approach can also be well generalized to other content analysis tasks.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Retrieval models*; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis –*Object recognition*

General Terms

Algorithms, Performance, Experimentation

Keywords

Affine Stable Characteristic, Sample Expansion, Object Detection

1. INTRODUCTION

Object can be the basic unit for multimedia content analysis, and often refers to user selected image region of interest. Aiming at establishing correspondences between similar objects that appear in different images, object detection plays an important role in many computer vision applications, such as image registration, copyright protection, commercial retrieval, photo management, and security surveillance [1, 2, 3].

Given a query image of an object, the objective of object detection in this paper is to determine whether or not the object is present in a large sample gallery. And, if present, return the corresponding source image to user. This task is very challenging because of the infinite geometric and photometric transformations such as blurring, cropping, adding noise and varying illumination condition. Among them, the changes of camera viewpoint and distance are the most difficult to deal with, because they induce apparent deformation of the appearance and scale of objects [4, 5, 6], which make it very hard to extract invariant features from them.

The essence of object detection is building feature model for each object in the dataset, and at the query time, establishing the correspondences between an unknown object and its most similar feature model [5, 6]. Given a large enough set of object samples taking from different viewpoint and distance, we can model any object as accurately as we can, but it is often not the case in practice. For some applications such as copyright protection and security surveillance, object samples in corpus are often very limited, while query images can be taken from various viewpoint and distance. An extreme case is to model object variations from only one sample, which is the main point we are concerned about in this paper. As shown in Figure 1, an on-line query object is marked with yellow rectangle, and the detection will performed on a large dataset, and there is only one sample for each object in this sample gallery.

To refine object model within limited training samples, sample expansion is introduced and has been successfully applied [6, 7, 8, 9, 10]. Generally speaking, traditional methods can be classified into two categories: on-line and off-line sample expansion.

On-line automatic sample expansion is a kind of passive processing [1, 6, 7]. Given a query object selected by user, the system extracts an original object feature model from this query region, and find out similar images from sample gallery. Those highly ranked images from the original query are used to refine the model and reissue a new query. This process can be repeated

CIVR'10, July 5–7, 2010, Xi'an, China.
Copyright 2010 ACM 978-1-4503-0117-6/10/07…$10.00.

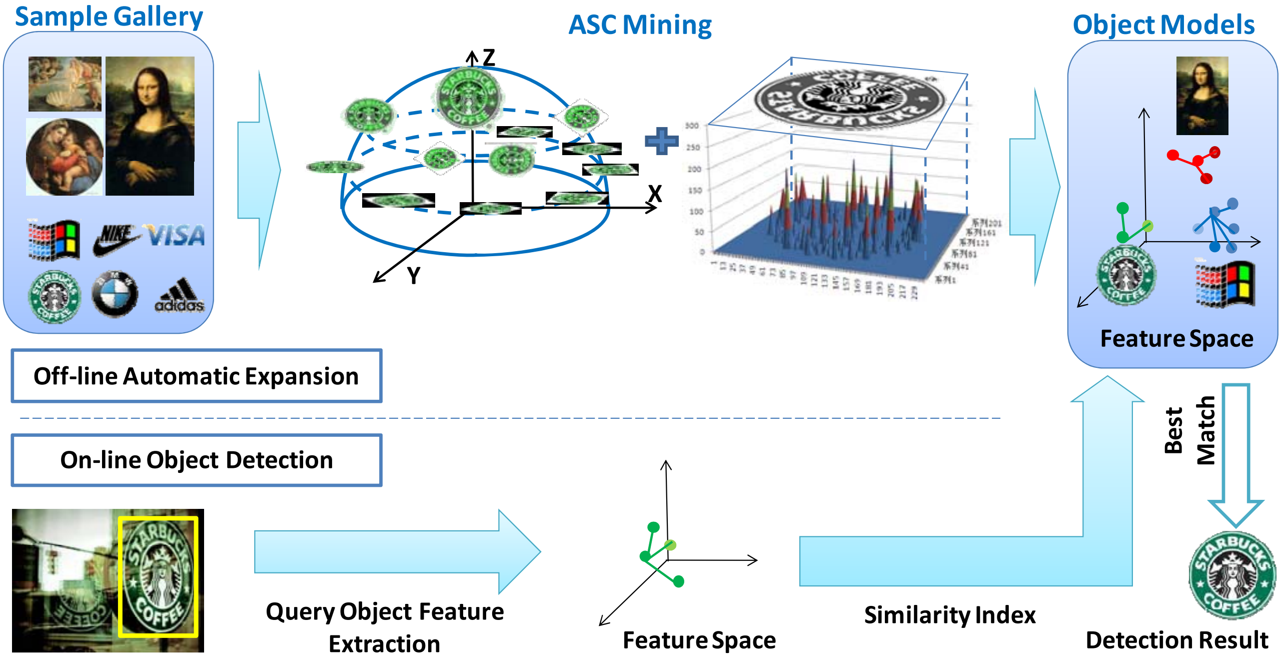

Figure 1. System architecture for object detection using Affine Stable Characteristic.

for several times. These richer generative models are expected to achieve better retrieval performance than the initial one. However, this kind of blind relevance feedback method has three disadvantages: (1) its performance relies heavily on the property of image dataset. If there are no enough object samples within different viewpoints and scales, the improvement will be very limited. (2) It can fail if false samples (not containing the query object) are included in the initial result set, because the new generative model will be incorrect and leading to divergence as the process is iterated. (3) On-line operation will protract detection time badly especially when there is a large sample gallery and several iterations are needed to get satisfactory result.

Off-line expansion method can be looked as a kind of active learning technology bloomed in recent years [8, 9, 10]. For each sample in the gallery, several synthesized images are generated automatically based on transformation simulation, and then a richer model can be learned from them. It is very necessary and useful for some specific applications, and is also the basis idea of our approach. The pioneer work suggesting simulating affine deformations can be traced to [10], and recent work in [8, 9] adopted similar methods. However, they often use empirical parameters which are neither comprehensive nor precise enough. To the best of our knowledge, ASIFT method proposed in [11] is the first systematic framework to simulate all representative image views. The creative work has received impressive performance in two image matching application. However, as to object detection in large corpus, this algorithm doesn't work well because of the following reasons: ASIFT simply combines all features obtained from synthesized images without selection. It will introduce lots of redundant features which not only result in false matches, but also consume memory and detection time.

We believe both robustness and distinctiveness are crucial for object feature model, and only those unique characteristics for each object should be retained. As shown is section 3, by integrating affine simulation with stable feature mining, we can obtain a compact and informative object model from only one object sample. The contributions of this paper are:

- We propose a new automatic off-line sample expansion approach suitable for object detection in large-scale dataset, and requiring only one object sample. The property is very important for some applications with limited training data.
- In order to obtain an object feature model with high robustness and distinctiveness simultaneously, we expand [11] by integrating a novel mining method: Based on stability calculation in two different but complementary hierarchies, the most stable SIFT features existing in adjacent simulated images are explored. They are regarded as unique characteristics for each object, and then compose a compact and informative object model called Affine Stable Characteristic (ASC for short).
- As this is an off-line expansion operation, and the dataset index has been reduced in the mining process, so applying our ASC method in object detection will not only increase the robustness and distinctiveness of object model, but also guarantee on-line detection speed.

The rest of this paper is organized as follows. Section 2 briefly reviews the state-of-the-art invariant local feature, explaining why we choose SIFT detector and descriptor as our basic method. Our ASC algorithm is described in section 3. Finally section 4 present extensive experimental results, and section 5 conclusions the major findings in this paper.

2. RELATED WORK

To achieve precise object detection result from a large gallery with limited object samples, the robustness and distinctiveness of visual feature are both very importance. Due to the invariance property for various geometric and photometric transformations, local image detectors and descriptors [12, 13, 14, 15, 16] has achieved brilliant success for object based applications [17, 18, 19]. However, none of them is fully affine invariant [11, 12, 16], so the object would be poorly represented by limited training data, not to speak of by only one sample.

Generally, a solid plane object's apparent deformation arising from a change in the camera position can be locally modeled by affine planar transforms [8, 9, 10, 11]. In the well-known work of [12, 13], they compare the performance of six famous affine covariant region detectors (not including SIFT algorithm [16] which is designed only to be scale invariant) under varying imaging conditions. Among them, MSER [14] and Hessian-Affine [15] have been demonstrated to have better performance than other detectors. However, none of these detectors are fully affine invariant, as they start with initial feature scales and locations selected in a non-affine-invariant manner, and only a small fraction of features can be matched for viewpoint changes beyond 30 degree. Moreover, when a strong change of scale is present, in practice larger than 3, SIFT detector still beats all other methods [11, 16].

SIFT algorithm [16] combines a scale invariant region detector and a gradient based descriptor. By simulating the zoom with different Gaussian convolutions, SIFT detects points of interest at extremum of scale-space, and sample for each extrema a square image patch, whose x-direction is the dominant gradients around the point. Moreover, the SIFT descriptor is represented by a 3D histogram of gradient locations and orientations, which makes the descriptor robust to small geometric distortions and small errors in the region detection. SIFT has been confirmed to be the only fully scale invariant method [11, 13].

In order to make SIFT fully affine invariant, [11] proposed a systematic framework to simulate all representative camera views, which has been approved very useful in two image matching application. However, when the proposed method is applied to object detection in a large corpus, it will introduce lots of redundant features which not only result in false matches, but also consume memory and detection time. Consequently, we expand this approach by combining a novel mining method to obtain an object feature model with both high robustness and distinctiveness.

3. AFFINE STABLE CHARACTERISTIC

3.1 System Overview

The general approach for our ASC based object detection framework is illustrated in Figure 1. Firstly, given an object sample, we perform off-line automatic expansion and feature mining to find the most stable and distinctive characteristics. An index for all of the model features contained in the sample gallery will be created. At query time, when user presents a query in the form of an image region, features are extracted without any on-line expansion and mapped into the index. According to the result of similarity searching, the image with the highest score which is also bigger than threshold will be regarded as the source image.

It is well-known that any planar smooth deformation of a solid object can be locally approximated by an affine transforms [8, 9, 10, 11]. In short, all local perspective effects can be modeled by local affine transforms: u (x, y) → u (ax + by +e, cx + dy + f) in image u (x, y). Among these six freedom parameters a~f, SIFT [16] is fully invariant with respect to four of them, namely zoom (scale), rotation and translation, while the angle parameters of camera axis have been left over.

Due to its almost perfect scale invariant property, we adopt SIFT to detect and describe local features. In order to make it fully affine invariant, we follow the approach in [11] to obtain some expanded samples as the first step of ASC. The details of our algorithm will be explained in the following subsections.

3.2 Affine SIFT Expansion

Follows the Singular Value Decomposition principle in matrix decomposition, any affine map with positive determinant has a unique decomposition:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \lambda \begin{bmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \qquad (1)$$

Where λ corresponding to camera zoom, ψ denotes camera spin. The parameter *t* is determined by θ, while angle θ and ϕ angles are respectively the camera optical axis longitude and latitude.

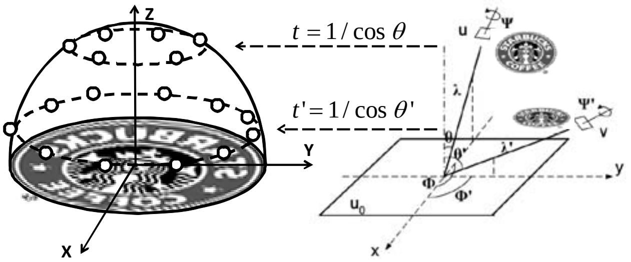

Figure 2. Geometric interpretation of the decomposition and influence of the camera optical axis.

Figure 2 shows a camera motion interpretation of the camera optical axis longitude and latitude, which result in apparent affine transformation and have been ignored by SIFT algorithm. Following the affine-space sampling method proposed in [11], a large set of synthetic images are generated by simulating the transition of camera optical axis on an observation hemisphere. Among these parameters, t is called the *tilt* and corresponds to simulated latitude angle. The range of t not only determines the amount of longitude sampling in each latitude circle, but also decides the total number of synthetic images. In [11], the suggested t = 1, a, a^2… a^N (a = $\sqrt{2}$, N is the number of latitude angle sampling), and the sampling step of longitude follow for each t: 72°/t. That is to say, when the number of tilt sampling N goes up to 5(corresponding to latitude angle of 80°), 41 synthetic images will be generated for each object.

On average, there are 3,300 SIFT local patches detected on an image of size 1024×768 [1, 16], and the number of descriptors is even larger because some patches have several dominant orientations. Even with a coarse sampling in a smaller range of latitude angle, the affine expansion will still result in an increase in the size of the feature database at least by a factor of 20.

3.3 ASC Mining with Global Stability

The expanded information indeed provides better robustness for affine changes, but also introduces a lot of redundant features which will not only need a much higher computational cost, but also result in a mass of false matches. To detect the most stable and useful subset of those features, and establish an object model with both robustness and distinctiveness, we proposed a notion called *Global Stability* to describe the importance of each feature detected in all synthetic images.

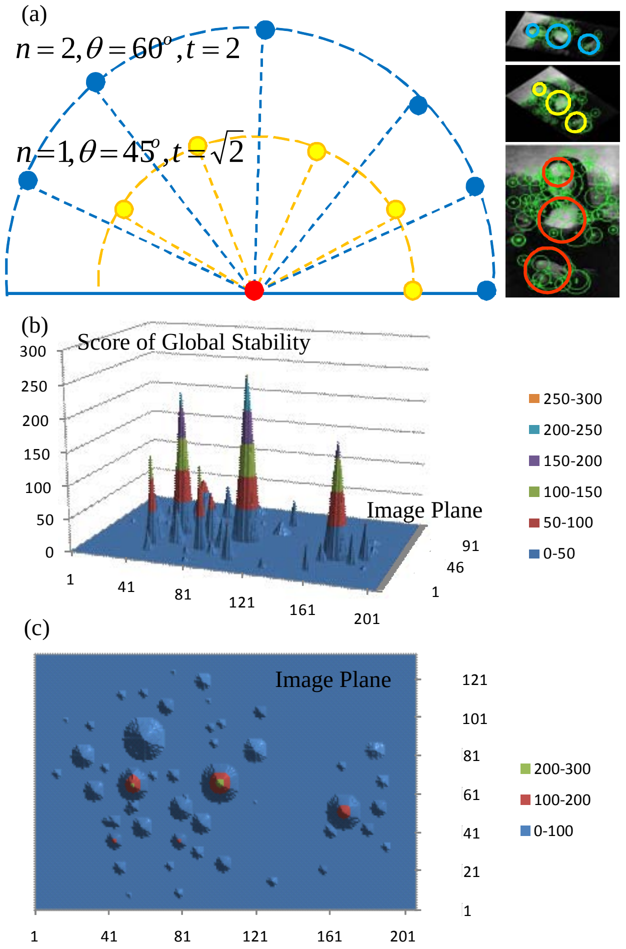

Figure 3. (a) Global Stability calculation and some prominent features (different colors denote latitude angles of 0°,45°,60°). (b)(c) Global Stability distribution of image Mona Lisa (perspective view and top view separately).

Although SIFT feature is designed to be scale invariant, but the combination of scale-space based detector and gradient based descriptor makes it also surprisingly invariant to small change of viewpoint (up to about 30 degrees). As shown in Figure 3 taking the Mona Lisa for instance, some stable features do exist even with wide viewpoint changes, such as local regions detected from her forehead, neck and hand (indicated by colorful circles). These SIFT descriptors extracted from synthetic images are very similar, thus can be matched with each other correctly. This shows that if a feature is repeatedly detected under different affine simulations, it is very likely to be useful for recognition and matching tasks.

Based on the above observation, we calculate global stability for each SIFT feature detected in the original object sample by combining feature matching and voting. The original sample is chosen as the base of feature matching, because it often has larger resolution and more details than synthetic images. The outline of this process is as follows:

(1) For each object sample *I* in the gallery, with the number of latitude sampling set to 5 (N=5), 41 synthetic samples will be expanded for each object, called *Affine-Space*. Extract SIFT features from the original sample and these synthetic images.

(2) For a simulation image *I'* in the affine-space of *I*, supposing it is obtained by affine mapping *H*, find matching features between *I'* and *I* based on Euclidean distance. It should be noted that two features are called "match" if and only if their coordinate and descriptors are both similar, and the coordinate of feature from *I'* will be projected back to *I* using $H^{-1}$for comparison.

(3) Repeat the above process to establish relation between *I* and every simulation image *I'*. For each pair of matching features, add score to corresponding features in *I*. And the total score for each feature is called *Global Stability* in this paper. To reduce the risk of feature location deviation, weights should be added according to Gaussian distribution, an example of voting result for an image is shown in Figure 3(b) (c).

(4) Finally, the score for each feature in *I* is normalized to the range of [0~255], which means the maximum score has been normalized to 255.

As illustrated in Figure 3(b) (c), peaks of score distribution denote local image patches with largest global stability, and corresponding features can be regarded as unique characteristics of this object. For simplicity, we only add score to corresponding features without Gaussian-weighted voting.

3.4 ASC Mining with Local Stability

Using global stability filtration, many redundant features can be largely reduced, and only the one with high distinctiveness will be reserved. However, take robustness into consideration, just considering global stability is far from enough. Because many features just be stable within limited affine transitions. These features are also very important for robustness and completeness of object model, especially when query image is taken from a large viewpoint change with the object sample. But they are very likely to be rejected with global stability filtration.

Consequently, we also proposed another notion called *Local Stability* to solve this problem. Different from global stability mentioned above, local stability is calculated for each image in the affine-space within a limited range of affine transitions. As illustrated in Figure 4, for each synthetic image *I'* expanded from *I*, a set of adjacent images are called its *neighbors* in affine-space. Supposing *I'* is generated with latitude angle θ_n and longitude angle ϕ_k, the images generated with $\theta_{n\pm i}$ and $\phi_{k\pm i}$ are called the *i-neighbouring* images of *I'*. The parameter *i* is set to 1 in this paper for simplicity. Given a synthetic image *I'* and its *i*-neighbouring images on the observation hemisphere, the calculation of local stability is similar to that of global stability. The only difference between them is that the calculation are just processed between each image and its *i*-neighbours, and the coordinates of each pair of matching features must be projected back to the original image *I* for voting. If the feature hasn't present in *I*, it will also be recorded with its projected coordinates. Finally, all of the features with high local stability are retained to generate the final object model. According to experimental results, we find that about half of features composing of the final object model are reserved because of high local stability. They contribute significantly to the robustness of object model.

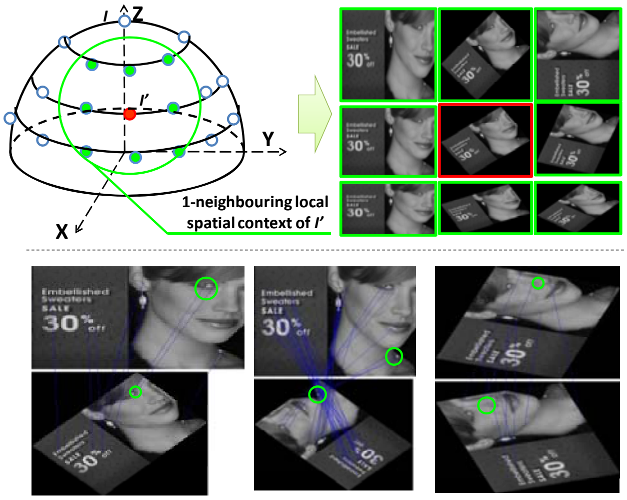

Figure 4. Illustration of 1-neighbouring Local Stability and some stable features newly found.

3.5 Object Model Generation

We adopt both global and local stability as criteria to evaluate the robustness of object features. They have potential relation but play different role in property description. In a word, global and local stability represent the robustness of each feature from different hierarchies, denote uniqueness and diversity separately.

Recall that our goal is to find those feature which maximizing the stability under wide range of affine simulations, so that a compact and informative object feature model can be generated for matching. Therefore, we propose to rank extracted features $\{f_i\}$ by using the combination of global stability and local stability:

$$object\ features\{f_i\} = \left\{ f_i \middle| \begin{array}{l} GS(f_i) > T_{GS} \\ or \\ LS(f_i) > T_{LS} \end{array} \right\} \qquad (2)$$

where $GS(f_i)$ and $LS(f_i)$ denote the global stability and local stability of feature f_i respectively, Here T_{GS} and T_{LS} are their thresholds, and the proper value of them will be determined experimentally in subsequent section. Because both $GS(f_i)$ and $LS(f_i)$ have been normalized to [0~255] according to the maximum score of each object, the value of T_G and T_L will also lies in this range, and are influenced by the score distribution of $GS(f_i)$ and $LS(f_i)$. Due to the variety of image content and resolution of different object, this way is more proper than choosing object characteristics with fixed numbers.

3.6 Object Detection with ASC

Given a query object, our task is to detect whether the object exists in the large sample gallery, and return the corresponding source image. The query object can be taken from arbitrary camera view and distance, which is very challenging for detection. Moreover, not only precision and recall, but also memory and time cost must be considered for practical applications.

Based the method introduced above, the most affine stable features called ASC are obtained for each object in the gallery, and their corresponding stability weights are also recorded in a dataset. To avoid exhausted search, we adopt the famous and practical method Approximate Nearest Neighbor [20] to establish index for quick feature matching. At query time, given a query in the form of an image region, we apply primary SIFT method [16] to extract query features $\{q_i\}$ without any expansion. For each query feature q_i, the top-K most similar features are found with L_2 distance, and K is set to 3. By counting the total sum of weight, object sample with the highest score and larger than 1.5 times the second-largest one will be regarded as the source image.

Although most of query features can find their correct matches in the large dataset, there also exist several false matches due to ground clutter. They can be further filtered by identifying their geometry consistency, and feature matches agree on the geometry consistency can be identified as correct with high confidence. This technology is very useful but not the main point of our discussion, and details can be found in [4, 17, 18, 21].

4. EXPERIMENTS

To evaluate our system, we perform our experiments mainly on two famous annotated dataset: 1) The *Oxford* dataset used in [1, 4, 18], this is a standard test set of 5,062 images with an extensive associated ground truth. 2) Another labeled dataset available from [17] is called *INRIA Holiday*, which contains 1,491 personal holiday photos. Moreover, to evaluate both the robustness and distinctiveness of our object models in actual applications, we also take more than 2,000 images chosen from video keyframes provided by *TRECVID* [22], and images crawled from *Google* which contain the most commonly searched objects, such as logos, landmarks and famous paintings.

4.1 Image Dataset

Datasets mentioned above contain a large variety of object types, more than 800 image groups in all. Most of the existing works used the first image of each group as the query image, and the other images of the group were regarded as the correct retrieval results. However, this method is not suitable for our approach, because what we are interested is object detection with very limited training data, which means that there is only one sample per object is available in our gallery. Consequently, we use the dataset in opposite way. Only one image for each object group was selected to form the sample gallery, and the rest are used as query images. Detection is regarded as successful if and only if the determination and matched object sample are both correct. What should be noted is that we have resized the samples in gallery to about 352×240 pixels due to computation concern. Apparently, with more object samples and higher image resolution, our detection performance should be even better.

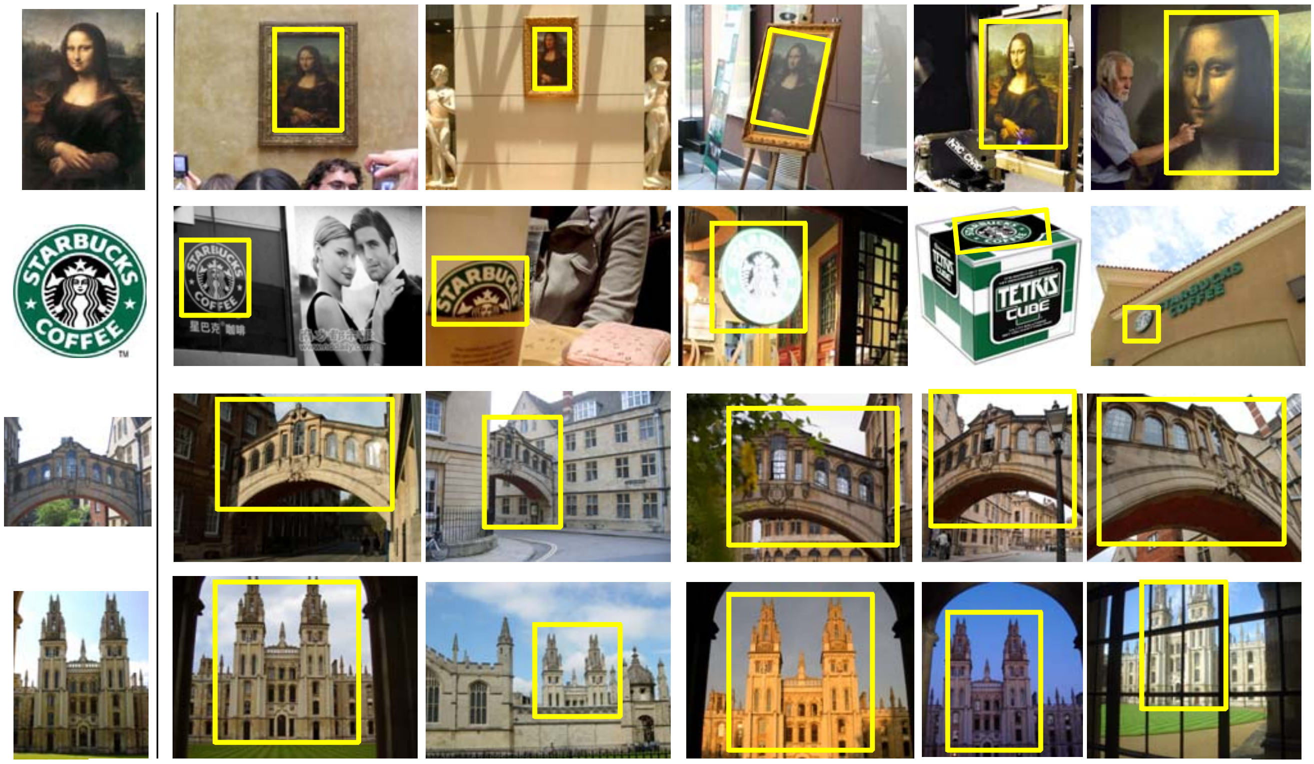

Figure 5. Some samples of results returned by ASC based detection (Mona Lisa, Starbucks, Hertford, All Souls).

For preliminary comparison, we first present the sizes of feature datasets extracted from the galley using different methods: SIFT [16], ASIFT [11], our ASC. The size of ASIFT's index is too big to be generated in practice, so the value is estimated by sampling method, which is accurate enough for comparison. In Table 1, we can see that the number of features and the size of index are both largely reduced by our method compared to ASIFT [11]. As shown in the following subsection, with the small feature set comparable with traditional methods [16, 19], we can achieve much better detection performance than them. This demonstrates that a lot of redundant features have been filtered successfully by our ASC, and only those stable and distinctive ones are retained.

Table 1. Size comparison of feature datasets and indexes

	SIFT[16]	ASIFT[11]	Our ASC
Number of features	496K	14,858K	697K
Size of ANN index	1,089M	33,628M	1,627M

4.2 Performance of Object Detection

For performance evaluation, we use precision, recall and F_1 similar with [1, 9]. Precision is the number of detected positive images relative to the total number of images detected. Recall is the number of detected positive images relative to the total number of positives in the corpus. An ideal precision-recall curve has precision 1 over all recall levels [1]. Here F_1 is define as $2pr/(p+r)$, where p and r indicate precision and recall respectively. A method with higher F_1 implies that its comprehensive performance is better than others.

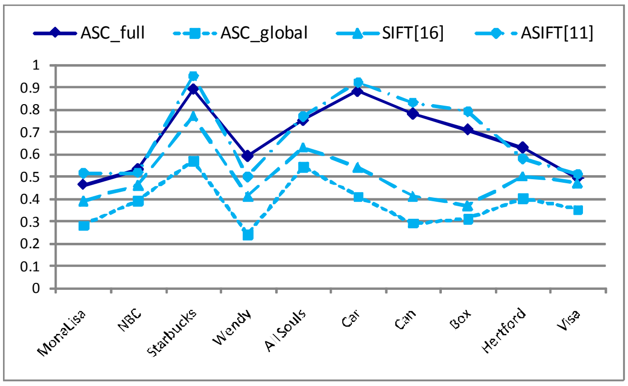

Figure 6. Performance comparison in F_1 measure.

Query images chosen for test all contain various transformations: rotations, viewpoint and illumination changes, blurring, etc. Some samples of challenging results returned by our method are shown in Figure 5(Objects are indicated by yellow rectangles). Due to the limit of space, we only present the performance of 10 object types in Figure 6. Their F_1 measures are calculated using four detection algorithms: SIFT [16], ASIFT [11], our method only uses Global Stability filtration (called ASC_global for short), and the full model that uses both global and local stability (ASC_full). Words in the horizontal axis denote object types. We don't choose other methods [1, 9] because they would degenerate to SIFT or ASC_global with only one sample and limited affine simulations. We didn't use original ASIFT as [11] because it is too memory-and-time-consuming to be applied in large-scale object detection, thus we only expand they query image using ASIFT method. Overall, the average F_1 increases 35.8% from 0.495 to 0.672 compared with SIFT. By combining both global and local stability, our ASC object model achieves the best performance.

As we can see, these methods consistently perform better on some object types than others. Best performances appear for textured objects such as Starbucks and Car, mainly because their abundant details are very suitable for local features. However, our method doesn't work well for texture-less objects such as Mona Lisa and Wendy, due to the small amount of salient local features. Figure 7 depicts the detection performance of Starbucks and Mona Lisa separately, using precision-recall curves of ASC.

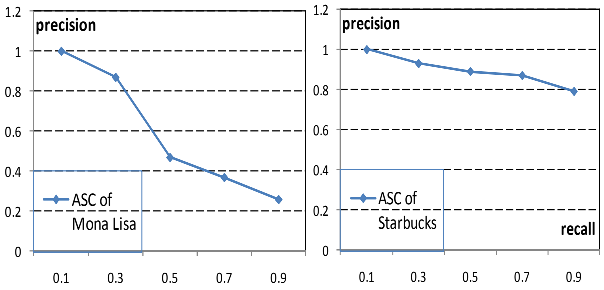

Figure 7. Precision-recall curves of objects with the best and worst performance respectively.

Another observation shown by Figure 6 is that the combination of global and local stability is very important for the robustness and distinctiveness of object feature model. That is to say, it is vitally important to choose the appropriate value of filtration threshold T_{GS} and T_{LS}. Therefore, to further examine the proposed algorithms, we also conducted experiments to study the influence of T_{GS} and T_{LS}.

Because both global and local stability are normalized to [0,255] in this paper, so the thresholds T_{GS} and T_{LS} will also vary in this range. Clearly, it is unpractical to compute the detection performance for all possible combinations of these two thresholds. With little loss of precision, we first find out the most suitable T_{GS} using F_1 measure, and then find T_{LS} with this T_{GS}.

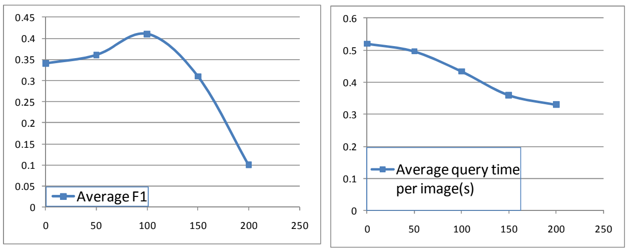

Figure 8. The effect of different T_{GS} on F_1 and query time.

We can see from Figure 8 that in our dataset the best performance appears when T_{GS} = 100, which can be explained as follows: when T_{GS} is very small, there are too many redundant features which would result in false matches and influence precision. But a large T_{GS} will reserve only a few features, and ruin the robustness and integrality of object feature model. Consequently, we adopt the value 100 throughout our experiment.

As illuminated before, combining local stability is very important for our ASC method. Hence we further examine the proper value of T_{LS} while T_{GS} = 100. It might seem surprising that the accuracy does not improve rapidly as more features are retained, as shown in Figure 9. The reason is that the decrease of T_{LS} indeed increase the robustness of object model by introducing many more features, but these features are on average less stable and distinctive, thus are less likely to be detected in the transformed query image.

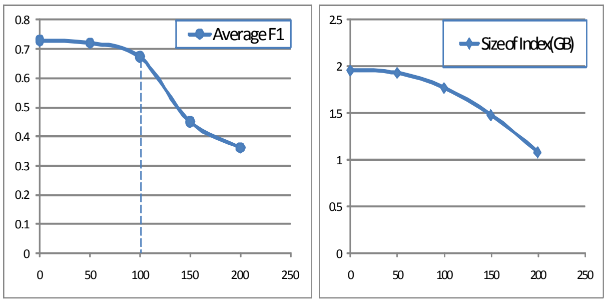

Figure 9. The effect of different T_{LS} on F_1 and index size.

Moreover, as illustrated by the second graph in Figure 9, size of dataset index will increase along with the number of retained features. Although the success of object detection often depends on the quantity of correctly matched features, the cost of computation and memory should also be taken into account. Accordingly, for the experiments in this paper we use T_{LS} =T_{GS} =100. We believe they are suitable to strike the right balance between performance and computation cost. And the size of index can be further reduced by dimension reduction methods such as [4, 17, 18].

4.3 Detection Speed

For object detection, the speed is also very important. As shown by the second graph in Figure 8 (the time cost of using ASC_full is only a little larger than ASC_global), the average query time per image of our ASC is about 0.43s. Compared to existing methods with similar precision but slower on-line speed [9], our ASC is more appropriate for practical applications. This is because that our method has transferred the learning process of automatic sample expansion from on-line to off-line, thus doesn't need any additional process at the query time, which would ensure the on-line detection speed.

During all steps of our ASC algorithm, the process of affine simulation is the most time-consuming, which is about 6.8s per image. The second one is stable feature mining, which will take 4.2s on average. As they are both offline operation, little effort has been devoted to optimizing efficiency in our experiments. But the approach described in this paper can be extended to incorporate any SIFT acceleration methods into the model representation. Besides, as GPU-based computation [23] has bloomed recently, and many feature extraction methods have been accelerated successfully to a time scale of milliseconds, we can adopt them to further improve our method.

5. CONCLUSION AND FUTURE WORK

Affine-invariant model generation with only one object sample is very important for many applications such as copyright protection and security surveillance. We develop a method for sample expansion-based affine stable object model generation, which has better detection performance compared to the state-of-the-art methods, while not sacrificing on-line detection speed. We propose a novel method called Affine Stable Characteristic to generate object model using only one sample. Two new notions, Global Stability and Local Stability, are introduced to calculate the robustness of each object feature from different hierarchies. By combining affine simulation and stable feature mining, a compact and informative object model which is robust to viewpoint and scale transformations is generated. Experiments show that Global Stability and Local Stability are both very important for ASC extraction. They have potential relation but play different role in property description. In a word, global and local stability represent the robustness of each feature from different hierarchies, denote uniqueness and diversity separately. The benefit of applying our ASC in scalable object detection, will not only guarantee on-line detection speed, but also increase the detection accuracy.

There is still much room to improve the approach. We expect that using more features such as color and spatial information would boost performance. Another future direction is combining the property of features with advanced index structure, to make our approach more suitable for large-scale applications.

6. ACKNOWLEDGMENTS

This work is supported by the National Basic Research Program of China (973Program, 2007CB311100); National High Technology and Research Development Program of China (863 Program, 2007AA01Z416); National Nature Science Foundation of China (60873165 60902090); Beijing New Star Project on Science & Technology(2007B071); Co-building Program of Beijing Municipal Education Commission.

7. REFERENCES

[1] O.Chum, J.Philbin, J.Sivic, M.Isard, A.Zisserman. Total recall: automatic query expansion with a generative feature model for object retrieval, *International Conference on Computer Vision*, 2007.

[2] O. Chum and A. Zisserman. An exemplar model for learning object classes, *Computer Vision and Pattern Recognition*, 2007.

[3] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid. Groups of adjacent contour segments for object detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:36–51, 2008.

[4] J.Philbin, O.Chum, M.Isard, J.Sivic and A.Zisserman. Object retrieval with large vocabularies and fast spatial matching, *Computer Vision and Pattern Recognition*, 2007.

[5] Y.G. Jiang, C.W. Ngo and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. *International Conference on Computer Vision*, 2007.

[6] Y.H.Kuo, K.T.Chen, C.H.Chiang, and W.H.Hsu. Query expansion for hash-based image object retrieval, *International Conference on Multimedia*, 2009.

[7] A.Holub, P. Perona, M.C.Burl. Entropy-based active learning for object recognition, *Computer Vision and Pattern Recognition Workshops*, 2008.

[8] J.H.Hsiao, C.S.Chen, L.F.Chien, and M.S.Chen. A new approach to image copy detection based on extended feature sets, *IEEE Transactions on Image Processing*, 16(8):2069–2079, 2007.

[9] W.Wu and J.Yang. Object fingerprints for content analysis with applications to street landmark localization. *ACM Multimedia*, 2008.

[10] D. Pritchard and W. Heidrich. Cloth motion capture. *Computer Graphics Forum*, 22(3):263–271, 2003.

[11] J.M. Morel and G.Yu, ASIFT: A new framework for fully affine invariant image comparison, *SIAM Journal on Imaging Sciences*, 2(2), 2009.

[12] K. Mikolajczyk, T. Tuytelaars, C. Schmid, etc. A comparison of affine region detectors, *In International Journal on Computer Vision*, 65(1/2):43-72, 2005.

[13] K.Mikolajczyk, and C.Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[14] J.Matas, O.Chum, M.Urban, and T.Pajdla. Robust wide-baseline stereo from maximally stable extremal regions, *British Machine Vision Conference*, pp. 384–393 , 2002.

[15] K.Mikolajczyk, and C.Schmid. Scale & affine invariant interest point detectors, *International Journal on Computer Vision*, 60(1):63–86, 2004.

[16] D. Lowe. Distinctive image features from scale invariant keypoints, *In International Journal on Computer Vision*, 60(2): 91- 110, 2004.

[17] H.Jegou, M.Douze and C.Schmid. Hamming embedding and weak geometry consistency for large scale image search, *European conference on Computer vision*, 2008.

[18] J.Sivic, A.Zisserman. Video google: a text retrieval approach to object matching in vdeos, *International Conference on Computer Vision*, 2003.

[19] H.Bay, T.Tuytelaars, L.V.Gool. SURF: speeded up robust features, *European Conference on Computer Vision*, 2006.

[20] S. Arya, T. Malamatos, and D. M. Mount. Space-time tradeoffs for approximate nearest neighbor searching, *Journal of the ACM*, 57: 1-54, 2009.

[21] K.Gao, S.L, Y.Z, S.T and D.Z. Logo detection based on spatial-spectral saliency and partial spatial context, *IEEE International Conference on Multimedia and Expo*, 2009.

[22] Http://www-nlpir.nist.gov/projects/trecvid/

[23] H.Xie, K.Gao, Y.Zhang, J. Li, Y.Liu. GPU-basd fast scale invariant interest point detector, *IEEE International Conference on Acoustics, Speech,and Signal Processing*, 2010.