Enabling Low Bitrate Mobile Visual Recognition – A Performance versus Bandwidth Evaluation

Yu-Chuan Su, Tzu-Hsuan Chiu, Yan-Ying Chen
Chun-Yen Yeh, Winston H. Hsu
National Taiwan University, Taipei, Taiwan

ABSTRACT

The rapid development of technologies in both hardware and software have made content-based multimedia services feasible on mobile devices such as smartphones and tablets; and the strong needs for mobile visual search and recognition have been emerging. While many real applications of visual recognition require a large scale recognition systems, the same technologies that support server-based scalable visual recognition may not be feasible on mobile devices due to the resource constraints. Although the client-server framework ensures the scalability, the real-time response subjects to the limitation on network bandwidth. Therefore, the main challenge for mobile visual recognition system should be the recognition bitrate, which is the amount of data transmission under the same recognition performance. For this work, we exploit and compare various strategies such as compact features, feature compression, feature signatures by hashing, image scaling, etc., to enable low bitrate mobile visual recognition. We argue that thumbnail image is a competitive candidate for low bitrate visual recognition because it carries multiple features at once and multi-feature fusion is important as the size of semantic space increases. Our evaluations on two subsets of ImageNet, both contain more than 10,000 images with 19 and 137 categories, verify the efficacy of thumbnail images. We further suggest a new strategy that combines single (local) feature signature and the thumbnail image, which achieves significant bitrate reduction from (average) 102,570 to 4,661 bytes with merely (overall) 10% performance degradation.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Object recognition*

Keywords

Mobile Image Recognition; Thumbnail Image; Bitrate; Multimodal Fusion

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2502110 .

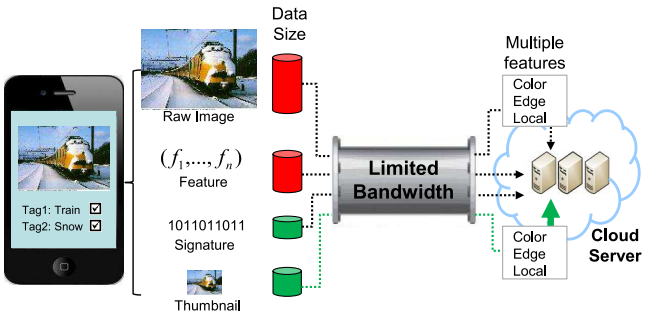

Figure 1: Low bitrate visual recognition. Images taken by a mobile device are classified on the server to ensure scalability over the large semantic space. The mobiles can send back information such as raw images, extracted features, compact signatures, or thumbnails, etc., for recognition in the server. Because the wireless connection between the server and mobile devices has limited bandwidth, the data transmission should be minimized to ensure prompt response, which is the main challenge for the mobile image recognition frameworks and also the key factor we would like to exploit in this work.

1. INTRODUCTION

Two ongoing trends are leading nowadays multimedia research and applications. The first is the increasing emphasis on the scale of data and the scalability of system. In visual recognition, the increase in scale does not only multiply the amount of data, but also the size of feature and semantic spaces [13, 28, 34]. The motivation for increasing the scale of visual recognition system is not only from research interests but also from real application needs. For example, since human can recognize tens of thousands of concepts and categorize them accordingly from images, an automatic photo annotation system would be limited if it can only recognize few concepts. Another example is event detection systems; since there exists an enormous amount of potential events, a general purpose event detection system would be applicable if more events can be detected in real-time response. Many new methods for scalable multimedia applications have been developed in the past few years [29, 34, 31], and the problem remains open with highly active research.

The second leading change is the paradigm shift from personal computer (PC) or workstations to mobile devices. While high quality camera becomes a basic component on current mobile devices, they are currently used as passive

recorders. Combining with the growing computation capability and the rich contexts in mobile devices, the camera can enable more proactive and smart applications such as remote healthcare, lifelog, automatic photo annotation, etc. Among all potential applications, many of them depend on visual recognition technologies. For example, automatic tagging system can be easily combined with popular social network applications such as *Facebook* to enable more user friendly media sharing.

1.1 Physical Constraints on Mobiles

Although there exists many promising technologies for large scale visual recognition, most of them assume operating on PCs or even workstations, where the underlying assumption about available resources may not fit that of mobile devices.

Some of the most important limitations on mobile devices include: **Limited computing power**, where many visual recognition systems assume the computing power of server level CPU. This restricts the use of complicated recognition systems and multiple features, where it takes roughly 1 second for feature extraction of each feature on current mobile devices [8, 10, 20]. **Limited storage**, usually around 10 to 100 GB. The limited storage inevitably restricts the amount of models that can be stored on the devices. Meanwhile, it imposes a hard limit on the scalability of purely native system. **Limited network bandwidth**, which forbids the rapid communication with remote servers. It also degrades the user experience when real time response is required for the services. **Limited power/battery**, which limits both heavy computation and rapid network transmission on the device. These limitations indicate that methods which work fine on servers may not work on mobile devices, and further optimization for mobile devices is necessary.

1.2 Low Bitrate Mobile Visual Recognition

With the physical limitations on mobile devices, the objective of mobile visual recognition system is to maximize user satisfactory under given resource constraints. User satisfactory is influenced by different factors such as response time, accuracy, power consumption and the scale of the semantic space. One possible direction for the mobile system is to match the server-based systems in terms of accuracy and scalability, which are the focus of server based system, while minimizing the response time and power consumption, which are the new requirements of mobile system.

For visual recognition on mobile devices, any native systems will have limited scalability due to the storage constraint. To overcome the constraint, the client-server framework is an intuitive solution and is adopted in many mobile visual search systems [9, 18]. The framework, however, is subject to the limitation of network bandwidth, as illustrated in fig. 1. We proposed the comparison of **recognition bitrate**, which considers the recognition performance with respect to the amount of data transmission between mobile devices and the server as in fig. 2. Unlike traditional server based systems, which consider only the dimension of performance, we introduce the new dimension of data transmission. The new dimension is important in that we would like to minimize the data transmission while retaining a reasonable performance in real visual recognition applications. Low recognition bitrate leads not only to faster response time for applications, but also lowers network usage rate

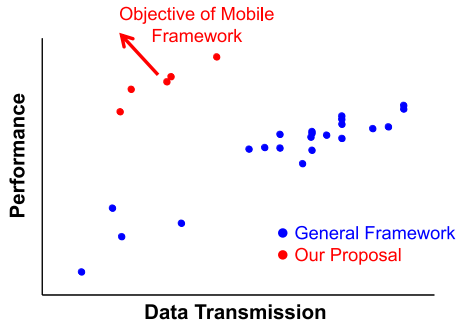

Figure 2: The goal for low bitrate mobile visual recognition. The objective of mobile visual recognition system is different from the general visual recognition systems, where the mobile system prefers the strategies that minimize the data transmission while ensuring the performance. For this work, we will exploit and compare various strategies such as state-of-the-art features, hashing learning methods, image scaling, etc., to enable low bitrate mobile visual recognition.

and battery consumption, which are important factors for real applications.

The limitation of wireless network on mobile devices has been widely acknowledged in mobile visual search [18]. For example, it takes 8 to 10 seconds on average to transmit a typical JPEG image over 3G network. Many efforts have been devoted to feature hashing or compression to enable visual search on mobile systems [33, 9], which provide some candidate strategies to achieve low bitrate mobile visual recognition including:

- *Transfer features with moderate dimension.* By using features with moderate data sizes and fair performance, the system can work with acceptable performance and data transmission.

- *Transfer compressed features.* The strategy exploits either feature with proper compression rate, such as compressed histogram of gradient, or compression schemes with small performance degrade, such as product quantization [29].

- *Transfer feature signature produced by hashing.* The strategy generates a compact signature that retains the performance. There exists a large family of hashing method, including random projection, spectral hashing, etc.

- *Transfer scaled-down images.* The images must be scaled down to be transfered over wireless network.

In most existing works on mobile visual search, transmitting images over the wireless network is considered infeasible for applications that require real-time response because of the image size. The claim, however, does not consider that it may not be necessary to transmit the original image to the server for visual recognition or retrieval. Research has shown that both human and computer system can recognize images with very small resolutions [31], which is much smaller in data size than general consumer photos. The fact that most existing visual recognition datasets are mainly

consist of small images also indicates that very high resolution images may not be necessary for visual recognition. Besides, mobile applications based on sending thumbnail images over wireless network have been proposed [11]. Indeed, while sending original images over 3G network may be impossible, transmitting thumbnail images is well applicable yet the efficacy of thumbnail images for visual recognition has not been explored.

The dire needs for mobile visual recognition emerge but the applicable methods are still missing. To entail mobile visual recognition, we conduct a systematic study on the recognition performance with respect to transmission bitrate for mobile visual recognition. In particular, we compare different strategies such as compact features, feature compression, feature signature by hashing, image scaling, etc. Several state-of-the-art features are included for comparisons. The evaluation is conducted on two subsets of ImageNet, both with more than 10,000 images. By understanding the bandwidth requirements and performance ranges for these strategies, we can motivate new mobile-cloud-balanced learning methods and cost-effective features. Our key contributions include:

- We propose to use recognition bitrate as the comparison criteria for mobile recognition models, instead of focusing only on recognition rate.
- We conduct intensive comparisons among various strategies for low bitrate mobile visual recognition, including visual features, feature signatures and thumbnail images. In particular, we examine the performance of thumbnail images with respect to image scales, which is not well studied in previous works.
- We combine multiple features in the comparisons of visual recognition bitrate. In existing works on low bitrate descriptor, the effort is on decreasing the bitrate of a single feature. Our experimental result shows that multi-feature fusion can further reduce bitrate and becomes more important when the number and diversity of concepts increase.
- We propose to transfer thumbnail image along with single (local) feature signature, which is affordable in both network bandwidth and computation for current mobile devices and achieves near optimal recognition performance.

The remaining of this paper is organized as follows. In section 2, we describe related works. In section 3, we describe the datasets used for the evaluation, and the general setup for the evaluation is in section 4. In section 5, we discuss the importance of multi-feature fusion. In section 6, we discuss the effect of image scaling on visual recognition performance. In section 7, we compare the recognition bitrate of signatures. Finally, we conclude our discovery in section 8.

2. RELATED WORK

One of the ongoing change in visual recognition research is the enlargement of semantic space. The most popular dataset for large scale visual recognition is ImageNet [12]. It is a dataset with concept ontology and contains more than 20,000 synsets, or concepts, with an average of 650 images for each concept. The ImageNet dataset has enabled the study of scalable visual recognition. Based on the dataset, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [4] is a contest that requires the recognition of 1,000 image categories. In ILSVRC 2010, there is a total of 1.2 million training images, which becomes one of the most popular image recognition dataset.

The development of visual feature has long been the center of visual recognition research. Among all visual features, local features have been proved effective in many domains. Many local features have been proposed since the advent of the difference of gaussian detector combined with SIFT descriptor introduced by Lowe [24, 25]; a large number of local features have adopted the SIFT descriptor with different detectors, such as Hessian affine detector and dense sampling [26]. For visual recognition systems, the local features from an image are usually pooled together to form a vector representation of the image for classification. There also exists many different pooling methods, such as Fisher vector or bag of visual word models [28, 30, 21]. Bag of word models can be further extended by multi-scaled spatial pyramid [22], which has been adopted in many state-of-the-art visual recognition systems.

While spatial pyramid method is superior in performance, it results in a high dimensional vector for each image and may even be larger than the original image in data size. The data size limits the scalability of systems and the transmission of features over wireless network in mobile devices. Feature compression methods such as product quantization have been introduced to increase the scalability of systems [29], where a compression rate of 64 to 128 can be achieved without significant loss of performance. Beside feature compression, various hashing methods that generate representative signatures from the original features are also studied, ranging from the data-independent random projection, data-dependent spectral hashing and semi-supervised sequential projection learning [5, 1, 33, 35].

With the explosive growth and prevalence of mobile devices, visual features that are more suitable for mobile devices have been developed. SURF is a local feature that reduce both computational cost and the dimension [3], and modification that makes the computation even more mobile friendly with moderate performance degrade is proposed [36]. Compressed histogram of gradient is a new local feature descriptor that aims to reduce the transmission bitrate [9], and residual enhanced visual vector is a new compact local feature pooling method that aims to reduce the data size and store the entire database for retrieval on the mobile devices [10].

Utilizing information in thumbnail images has been exploited before in different tasks. Torralba et. al. have conducted an object recognition study on a dataset of nearly 80 million tiny images with 32×32 pixel color images [31]. Their result shows that color images with 32×32 resolution are already recognizable by human. Thumbnail images are also used in new mobile services. In IMShare, a new mobile image sharing technique is built based on thumbnail images [11]. The thumbnail image of a photo taken by mobile devices and the detection results of local feature detector are sent to the server, where the server reconstructs the image by the thumbnail image and local feature descriptors extracted from the thumbnail image. These results show that thumbnail image do contain sufficient information for recognition, and that extracting meaningful features from thumbnail image is possible.

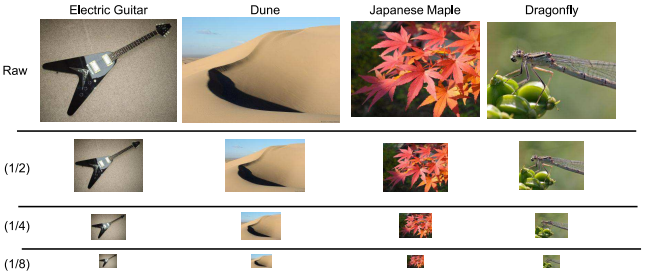

Figure 3: Example images from ImageNet137 dataset. The dataset consists of photos downloaded from internet. We also show the same images in different scales, as in the evaluation setting. Note the images are not shown in their original sizes but preserve their relative sizes and scales.

3. DATASETS

In this section, we describe the datasets used for the study. The goal of this paper is to study the recognition performance under different transmission bandwidths. In particular, we would like to study the performance of features extracted from different image sizes, which prohibit from directly using most of the popular datasets for visual recognition, because most of the datasets such as Scene [22] and Caltech256 [19] are mainly consist of images with low resolution compared with consumer photos. Besides, to evaluate the effect of increasing the scale of recognition system, the experiments require both large training set and semantic space.

To overcome the limit, we perform experiments on two subsets of ImageNet and screened out images with low resolution in the datasets. The choice of ImageNet is to ensure there exists enough images and classes for study. Specifically, we consider the following two dataset:

- **ImageNet19.** 19 categories from ImageNet 2011 Fall Release with the same categories of PASCAL Visual Object Classes (VOC) 2007 Challenge [15] were selected, following the protocol in [28]. We did not find the synset of "potted plant," and for the remaining 19 synsets, we downloaded only the synset itself without children synsets. In this dataset, images with length smaller than 500 pixels or width smaller than 300 pixels were discarded, which resulted in a total of 19886 images.

- **ImageNet137.** The training set of ILSVRC2010 was used to construct this dataset. In this dataset, images with length smaller than 800 pixels or width smaller than 600 pixels were discarded, and only categories with more than 60 images remaining were included, which resulted in 137 categories and a total of 12008 images. Some of the images are shown in fig. 3.

Note the two datasets have no overlapping categories. Following the convention of the PASCAL VOC Challenge [15, 16], each category was randomly and equally split into training and test sets, and we repeated the process 10 times. All results are averaged over the 10 runs of experiment with the statistical standard deviation also reported.

4. EXPERIMENTAL SETUP

In this section, we describe the features used in evaluation and the feature extraction procedures. Then we describe the classifier adopted for classification.

4.1 Features

To explore the optimal recognition performance given an image, we evaluate a variety of popular general-purpose features. Many of these features are adopted in the state-of-the-art visual recognition systems and are evaluated on both public datasets and recognition contests such as PASCAL VOC and ImageNet [16, 4, 28, 34]. The features can be categorized into the following two groups.

4.1.1 Global Features

Global features include all features that are non-local. While there exists a large number of global features ranging from color, texture to edge features, we choose some of the most widely used general-purpose feature as follow:

- **Color histogram**. We use 24 dimension color histogram in *HSV* color space where the histogram on each dimension is computed separately and then concated. For quantization, H is divided into 18 levels, S and V are divided into 3 levels.

- **Color Moment**. We use 225 dimension grid color moment; each image is divided into 5x5 grids and the first to third moments in *RGB* color space are extracted from each grid respectively.

- **Gabor**. We use 48 dimension log Gabor coefficients as features. 24 filters with 6 orientations and 4 scales are used to compute the response, and for each filter response, the first and second moment are extracted [23].

- **LBP**. We use local binary pattern with uniform patterns extension, which results in a 59 dimension histogram [2].

- **PHOG**. We use 3400 dimensional pyramid histogram of oriented gradient with 4 bin histogram and 3 level pyramid ($K = 4$, $L = 3$) [6].

We also examine the popular Gist [27, 14] feature, but the performance is only slightly better than color histogram on ImageNet19 despite of the high dimensionality (960), so we do not include it in further experiments.

4.1.2 Local Features

Local features have become the standard component of state-of-the-art visual recognition systems. Among the wide range of detectors as well as descriptors, we choose the following combinations in our evaluations:

- **DoG**. Difference of Gaussian detector + SIFT descriptor [24].

- **HA**. Hessian Affine detector + SIFT descriptor [26].

- **Dense**. Extract SIFT features with dense sampling, using 20×20 patches and overlapping windows shifted by 10 pixels. We use the Vlfeat library for Dense SIFT extraction [32].

- **SURF**. SURF detector + SURF descriptor [3].

We also examine Compressed Histogram of Gradient [9] (CHoG) descriptor which is especially designed for mobile visual search; we do not include it in the following discussions because its performance is similar to SURF in our preliminary test and is not especially representative.

4.2 Descriptors

To utilize local features in classification, local descriptors in an image are usually aggregated into a compact feature. Many descriptors were proposed in recent years, while Bag of Word (BoW) and Fisher Vector (FV) are the most popular ones among all descriptors. We choose the following two descriptors which are the variations of BoW and FV respectively in our evaluations:

- **LLC**. Locality constraint linear coding (LLC) is a variation of BoW and further combined with Spatial Pyramid (SPM). We choose LLC for its performance, and our preliminary tests also confirm that LLC significantly improves over BoW + SPM. In our experiments, we use codebook size $c = 200$ and 400 with pyramid level $l = 2$. Note that the codebook was constructed simply by K-means without optimization for LLC.

- **VLAD**. Vector of locally aggregated descriptors (VLAD) is a simplification of FV. The reasons for choosing VLAD over FV is twofold. The first is that in our preliminary tests, VLAD shows comparable performance with FV with smaller dimension (no covariance vector). The second is that computing VLAD requires less storage and computation resource, which is important on mobile devices. In our experiments, we use codebook size $c = 16, 64, 256$ respectively.

4.3 Feature Extraction

To evaluate the recognition performance with respect to varying image sizes, we extract the features from the same image at different scales. For ImageNet19 dataset, the images are scaled down to 1/2, 1/4, 1/8, and for ImageNet137 dataset, the images are scaled down to 1/2, 1/4, 1/8, 1/16, as shown in fig. 3. On feature extraction, every (scaled-down) image is scaled up to their original size using bilinear interpolation before feature extraction. Image up-scaling is performed for performance reasons; our evaluation shows that scaling up the images to the original size generally yields better performance. In particular, the image size directly corresponds to feature point number for Dense SIFT feature, which in turns affects performance. Therefore, we perform image up scaling for all features for the consistency of evaluation. No additional information other than the thumbnail image is used during feature extraction.

We also examined the performance of various feature extraction strategies. For example, we computed the SIFT descriptors on upscaled images using the salient points detected on the original images, which is the strategy adopted in *IMShare* [11]. The additional detector information in this strategy turns out to be unhelpful for recognition. We also examined the optimal scales up for thumbnail images on feature extraction, and the result is in favor of scaling up image to the original size.

4.4 Classifier

In this section, we describe the classifier adopted for classification and the method for multi-feature fusion. For all features, we use SVM with linear kernel for classification [17] with 1-vs-all framework for multi-class classification. Linear SVM is adopted because of its training and testing efficiency and its success in many state-of-the-art visual recognition system [16, 4, 28, 34]. The parameter of SVM is determined by 5-fold cross validation on the training set.

For multi-feature fusion, we apply late fusion strategy – averaging the normalized scores from different classifiers over varying features. We use late fusion instead of early fusion for its efficiency, which is important due to the large number of possible combinations. To perform late fusion, the decision values of each classifier are first normalized with sigmoid function, and the scores from the same modality over different categories of each image are then $L1$ normalized. The summation of normalized score over all features is used to determine the class label of the instance. We do not explore sophisticated fusion methods, because the focus of this study is on features rather than the sophisticated algorithms, which do not show significant performance gains.

4.5 Compression Factor of Images

For bitrate comparison, we have to estimate the image size (in storage). The average image size over the entire dataset is used as the estimator, and the images are all in JPEG format with the original image quality of ImageNet dataset. For thumbnail images, Image scaling is performed using OpenCV [7], and the thumbnail images are also in JPEG format with the default compression factor 95 in 100-scale of OpenCV.

Although changing the compression configuration may also reduce data size, we focus on image scaling because the efficacy of thumbnail image has been justified [31, 11]. Besides, image scaling is more straightforward for controlling image quality. Therefore, the compression configuration is kept the same throughout the evaluations.

5. MULTI-FEATURE FUSION IS IMPORTANT

We first evaluate the performance of different features and show that no single feature performs the best on all categories. The result implies the importance of multiple features, and we show that multi-feature fusion is more efficient than increasing feature and descriptor dimensions in improving image recognition performance. The overall classification performance of each feature for the two datasets are in fig. 4. The result shows that, on average, local features significantly outperform global features. Besides, Hessian affine SIFT performs better with VLAD, while Dense SIFT performs better with LLC and achieves the best performance when we consider only single feature.

Although the performance differences between different features seems significant, the situation is very different when we inspect closer about the classification results on category basis. The classification accuracy of 6 out of 19 categories in ImageNet19 are in fig. 5(A). We can see from the result that no single feature achieves the best performance in all categories, and in some categories, global features are comparable with local features. This implies the necessity of using multiple features to achieve robust overall classification performance.

The same observation can be made in ImageNet137 dataset, where the result of 12 out of 137 categories are in fig. 5(B). Note that nearly every feature, including global features, achieves the best in certain categories. The result leads to

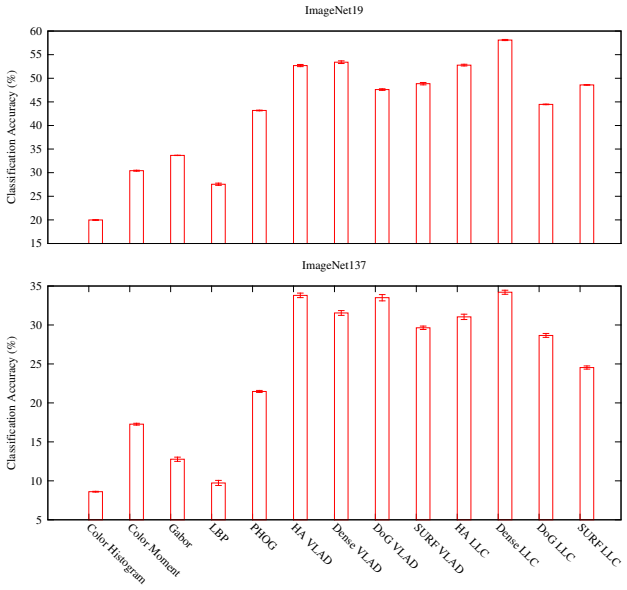

Figure 4: Average classification accuracies of different features on the original image. $c = 64$ for VLAD and $c = 400$ for LLC descriptor. On average, local features significantly outperform global features, and Dense SIFT + LLC achieves the optimal performance on both datasets. All results are reported with statistical standard deviation over 10 rounds of experiments.

the same conclusion that multi-feature fusion is important; it further indicates that multiple features are getting more important as the number and diversity of the categories to be recognized increases.

Based on the observation, we perform late fusion of multiple features for classification to verify the importance of multiple feature. We perform feature selection on late fusion by iteratively adding one feature at a time, where the feature that achieves the most performance improvement is selected. The process stops when no further features can improve the performance. The result is in fig. 6, where both absolute and relative improvements of each feature fused are reported. The relative improvement is defined by the absolute improvement divided by the optimal fusion performance. We can see that multi-feature fusion, even with a simple late fusion (i.e., averaging the normalized confidence scores from different modalities) strategy, can significantly increase the classification performance, and the relative improvement increases as the number of category increases. This result is consistent with the previous observation that the importance of multi-feature fusion increases as the diversity of categories increases.

A commonly used strategy to increase classification performance while using single feature, especially local features, is to increase the descriptor dimension. We next compare the effectiveness of increasing descriptor dimension and fusing multiple features. In particular, we compare the classification accuracy with respect to the total feature dimension,

[1]Since we have conducted intensive experiments of different configurations, all the figures are best seen in color.

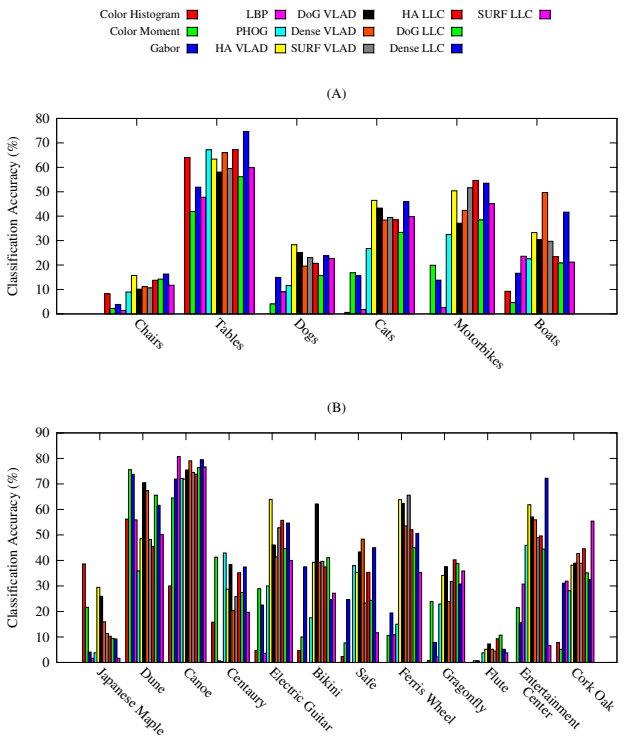

Figure 5: Results of 6 categories from ImageNet19 dataset are in (A), 12 categories from ImageNet137 dataset in (B). From ImageNet19, we can see that no single feature achieves the best performance in all categories. This indicates the necessity of multiple features to achieve optimal classification performance. In ImageNet137, which contains more categories and diverse concepts, every feature except Gabor achieves the best performances in different categories, and even the same local feature using different pooling methods show different performances. Compared with the results of ImageNet19 dataset, we can see the strong needs of multi-modal fusion across different features as the number of category increases[1].

because the dimension is proportional to the bitrate of the strategy. The dimension of multi-feature fusion is defined as the sum of the dimensions of all the features being fused. The result of ImageNet19 dataset is in fig. 7. It is obvious that multi-feature fusion achieves better performance than increasing the feature dimension under the same bitrate. The result indicates that when extracting multiple features is possible, using multiple features is more efficient for improving classification accuracy than using complicated models from a single feature.

6. IMAGE SCALING REDUCES BITRATE

In this section, we evaluate the recognition performance of features extracted from scaled down images and the recognition bitrate of the image. Our evaluation is based on the general claim that images taken by mobile devices are too large to be transfered over wireless network for prompt response time. In visual recognition, for example, transmitting images in the original high resolution may not be necessary,

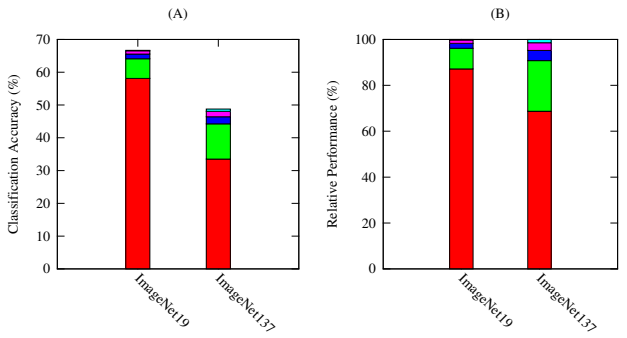

Figure 6: Results of multi-feature fusion. For (A), each section in the histogram indicates the absolute performance improvement of fusing one additional (best) feature. For (B), each section indicates the relative performance improvement over the optimal performance. The relative improvement increases in dataset with larger number and diversity of categories (ImageNet137), which confirms that multiple feature fusion is very important as the category (or concept) number increases.

because images with transferable size may achieve acceptable performance for applications.

Because the image size is dependent to the image content, we measure the average image size of the dataset for bitrate comparison. The result of ImageNet19 dataset is in table 1. Note that the average size difference between color image and gray scale image is small compared to the effect of image scaling; so we assume transmitting color image in the following discussions.

We first measure the performance degradation incurred by image down-scaling. The result of global features is in fig. 8, and that of local features is in fig. 9. We can see from the result that global features are more resistant to image down-scaling, while local features degrade more significantly. The performance of PHOG may even exceed DoG SIFT and SURF when the images are scaled down to 1/8 on ImageNet19 dataset. This indicates that multiple global features can be carried by a single scaled-down image without significant loss of information.

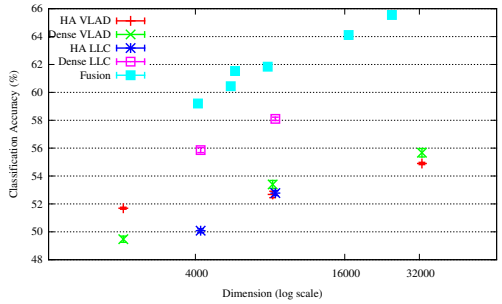

Figure 7: Recognition performance versus total dimensions of features on ImageNet19 dataset. Fusing multiple features, including global features, can significantly improve performance under the same feature dimension. Note that we only apply a simple late fusion strategy (i.e., average confident scores) to achieve the significant performance improvement.

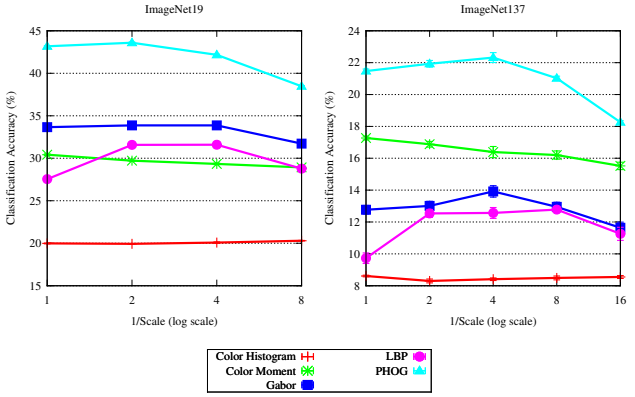

Figure 8: The performance changes of global features under different image scales on ImageNet19 dataset. The performance of global feature does not degrade significantly as the image being scaled down even to a tiny image. Therefore, we can compress and transmit all global features at once by sending a scaled down image without much loss of performance.

We next compare the performance of scaled-down image and image features under different bandwidth requirements. Based on the previous observation, we know that scaled down image contains information of multiple (global) features, therefore, the result of multi-feature fusion as well as that of the best single feature are reported. The result of ImageNet19 dataset is in fig. 10. The feature sizes are computed using feature dimension, where each dimension is stored in a double precision floating point number. We can see that performing multi-feature fusion with scaled down image achieves better performance under the same bandwidth, and the performance even does not degrade under moderate scaling. The result in ImageNet137 dataset is similar, as can be seen in fig. 11, with more significant improvements by multi-feature fusion.

The observation indicates that in mobile visual recognition, transmitting the high-resolution image over wireless

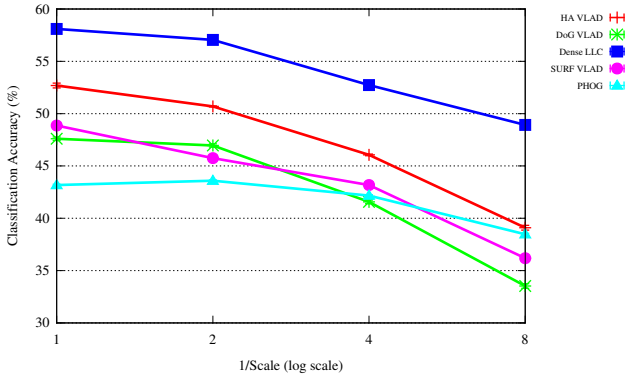

Figure 9: The performance change of local features under different image scales on ImageNet19 dataset. Local features degrade more significantly while image is scaled down. Note that the performance of PHOG even exceeds some local features when the image is scaled down to 1/8.

Table 1: Average image size of ImageNet19 dataset. The average size of color images is slightly larger than that of gray images but contains color information which is important in some categories. The trade-off supports transmitting color images over gray images, and we assume transmitting color image in all following discussion.

	SIFT LLC(400)	SIFT VLAD(64)	Original Color	1/2 size Color	1/2 size Gray	1/4 size Color	1/4 size Gray	1/8 size Color	1/8 size Gray
Bytes	67,200	65,536	102,570	31,581	27,094	10,352	8,711	3,637	2,912

network is impossible due to the original image size; meanwhile, it is not necessary to transmit the image in original size. Transmitting scaled down image may greatly reduce the bitrate without loss of performance; and by exploiting informations from multiple features, scaled down image may be more transmission efficient than image features.

Another benefit of transmitting scaled down image over transmitting features is the reduction of computation and storage overhead on mobile devices. Image features such as LLC require the storage of codebook and solving linear equations on the fly, and extracting multi-feature for real time application is computation intensive. Transmitting scaled-down image eliminates all the overhead, where the resource on server can handle these overhead without difficulty.

7. FEATURE SIGNATURE ACHIEVES LOWER BITRATE

In mobile visual search, much efforts have been devoted to generating a compact signature from the raw features to reduce the bitrate for retrieval. In this section, we compare the performance versus bitrate between scaled down image and image signature.

For signature generation, we use the data independent sparse random projection (RP) [1]. To verify the choice, we also compare the performance with two state-of-the-art hashing and compression methods such as the unsupervised product quantization (PQ) [29] and the supervised sequential projection learning hash (SPLH) [33], on ImageNet19 dataset. The result is in fig. 12. For PQ, the feature vector is first divided into subvector, with the dimension of subvector being $G = 8$ for VLAD and $G = 10$ for LLC. The average number of bits for representing each dimension is set to $b = 1$. For RP and SPLH, we set the projection matrix as a square matrix so the output bit number is the same as the input dimension; so the compression rate is 64 for all methods. Note the bit number of signatures are large (8k) to ensure recognition performance, because our goal is to build a mobile system with its performance comparable with server-based system. We can see from the result that SPLH does not perform as good as RP in high dimensional (8k) signatures; more importantly, it performs poorly with LLC. The performance of PQ and RP is comparable, and we choose RP for further experiments because of its computation efficiency and data independent property.

The result of signature is in fig. 13. We can see that the bandwidth requirement for signature is lower than that of the scaled-down image. This suggests that there exists redundant information in the scaled image which is not fully utilized by the recognition system yet. We also examine the performance of fusing multiple feature signatures. Under the two aspects of mobile visual recognition, that is, the recognition rate and bitrate, fusing multiple signatures turns out to be the best strategy with near optimal performance and roughly the same bitrate as thumbnail images. Note that the performance of fusing multiple feature signatures can not be further improve by including more signatures under our multi-modal fusion framework; it might indicate that there exists irreversible information lost in signature generation.

Although multiple (local) feature signatures achieve almost the best performance with low bandwidth, the strategy may turn out to be unfeasible when we consider the constraints in mobile computing. The most significant problem lies in the computing power on current mobile devices,

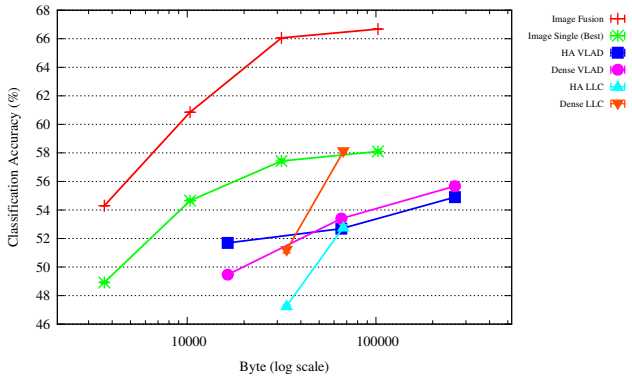

Figure 10: The recognition bitrate of scaled down image on ImageNet19 dataset. Because a scaled down image contains information of multiple features, the performance of both optimal single feature and multi-feature fusion is reported. With multi-feature fusion, scaled down image can significantly outperform the raw feature under given bitrate.

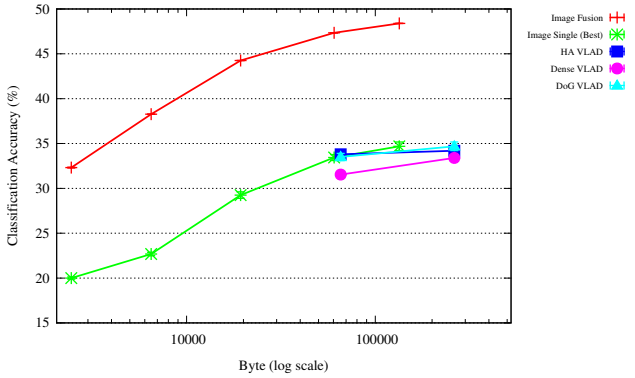

Figure 11: The recognition bitrate of scaled down image on ImageNet137 dataset. Although single feature performance degrades more significantly on ImageNet137 dataset, the fusion result of thumbnail image still outperforms raw features in recognition bitrate.

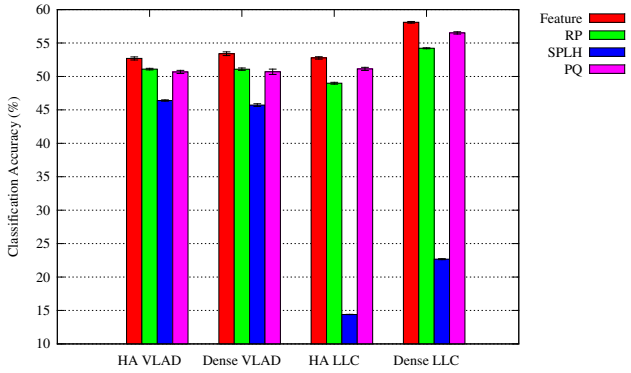

Figure 12: Performance comparison of different signature generation methods. Feature stands for the performance of the original feature. $c = 64$ is used for VLAD and $c = 400$ for LLC. For random projection (RP) and sequential projection learning hash (SPLH), the output bit number is the same as input dimension. For product quantization (PQ), the average bit number for each dimension is set to $b = 1$. Therefore, the compression rate is 64 for all methods. Note that we use high dimensional signatures (8k) to ensure that the recognition performance does not degrade significantly. Under the compression rate, RP outperforms SPLH, and SPLH performs poorly with LLC. The performance of RP and PQ is comparable; we use RP for its efficient computation in the following experiments.

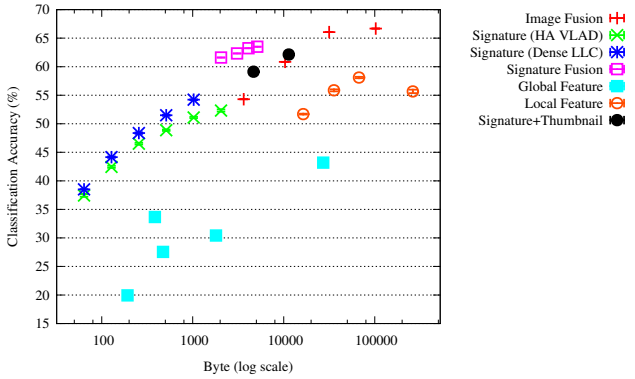

Figure 13: Bitrate comparison of all strategies, including feature signatures. Feature signature achieves lowest bitrate under similar accuracy. For the signature used to fuse with thumbnail image, we use HA+VLAD+RP because VLAD is more mobile friendly than LLC. Among all strategies, multiple feature signatures achieves lowest recognition bitrate, but is not feasible on current mobile devices due to the computation overhead of extracting multiple features on the devices. Combining single (local) feature signature and thumbnail image is the solution that best fits current mobile device constraints.

because it requires extraction of multiple features solely on the device which is computation intensive. Fortunately, it is feasible to compute at least single feature signature on mobile devices which is the basis of many mobile visual search system. Based on our own implementation, it takes less than a second on average to compute the signature of VLAD with SURF feature using codebook size $c = 64$ on iPhone 5. Therefore, a more realistic solution is to compute single feature signature on the device and send both the thumbnail image and feature signature to the remote server. The result of fusing thumbnail images and single feature signature is in fig. 13. We use Hessian affine local feature and VLAD descriptor with $c = 64$, with the 8,192 bits signature generated by sparse random projection. Note that we do not choose the signature with optimal performance (Dense+LLC+RP) because LLC is computation intensive and is formidable for mobile environments. The strategy balances among different constraints on the device, i.e., storage, network bandwidth and CPU, and it achieves almost the best performance with moderate bitrate. Under current physical constraints on mobile devices, this is probably the best strategy from our evaluations in terms of recognition bandwidth requirements.

8. CONCLUSIONS AND FUTURE WORKS

This paper presents a study on possible solutions for scalable mobile visual recognition system. We conduct a systematic evaluation on various strategies for mobile visual recognition under client-server framework in terms of recognition bitrate, and the experiments are conducted on two subsets of ImageNet; both datasets contain more than 10,000 images and one of them contains more than 100 categories. In particular, we compare the recognition bitrate of thumbnail images, various image features and feature signatures. The result shows that even a tiny image contains sufficient information for visual recognition, and by utilizing multiple features extracted from thumbnail images, the recognition bitrate of thumbnail image is much lower than raw image features. Although fusing multiple image signatures may achieve lower bitrate than thumbnail images, extracting multiple (local) features on mobile devices may not be feasible due to CPU and battery constraints. These experiments indicate that transmitting thumbnail images should be considered for mobile visual recognition systems.

We further recommend to combine single local feature signature and the scaled-down thumbnail image, which achieves near optimal performance under the constraint of current mobile environment. Using the strategy, we significantly reduce the average data transmission from 102,570 bytes to 4,661 bytes (i.e., thumbnail images scaled down to 1/8 of the original size and the 8,192 bits signature generated by random projection from Hessian affine feature with 64 centers VLAD descriptor), while the recognition accuracy only decreases from 0.67 to 0.59, which is still better than any single raw feature.

With the new aspects introduced in mobile visual recognition system, such as network bandwidth, power consumption, etc., there remains many to be explored. In this paper, we focus on the dimensions of feature and data transmission, and we will extend into other aspects in the future studies. In particular, we would like to extend the evaluation to real system and photos taken by mobile devices in the future, where the real world environment may bring new challenges. We would also like to evaluate the recognition bitrate of videos, because videos are generally larger in data size and consume a significant amount of mobile network bandwidth, and the evaluation on videos is expected

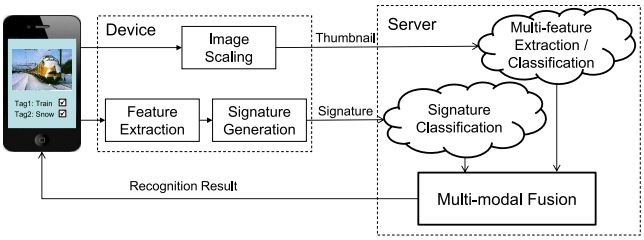

Figure 14: A recommended system design based on our evaluation results. Single (local) feature signature (e.g. HA+VLAD+RP) is computed on the mobile device, and is transmitted back to server along with the (down-scaled) thumbnail image. Server extracts multiple features from the thumbnail image and perform multi-modal fusion, then returns the results to the mobile device.

to have an even larger impact. Other exciting directions include better signature generations, where the side information from thumbnail images may be used to improve the compression rate and achieve lower bitrate. Similarly, a better compression technique for images may also lead to better recognition bitrate. Beside data compression on either thumbnail images or signatures, a better classification algorithm on the signature is also possible, where the signatures lie in Hamming space rather than Euclidean space in which general classification algorithms are developed.

9. REFERENCES

[1] D. Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *J. Comput. Sys. Sci.*, 66, 2003.

[2] T. Ahonen et al. Face recognition with local binary patterns. In *ECCV*, 2004.

[3] H. Bay et al. Surf: Speeded up robust features. In *ECCV*, 2006.

[4] A. Berg et al. Large scale visual recognition challenge 2010, 2010.

[5] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *ACM SIGKDD*, 2001.

[6] A. Bosch et al. Representing shape with a spatial pyramid kernel. In *ACM CIVR*, 2007.

[7] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[8] V. Chandrasekhar et al. Comparison of local feature descriptors for mobile visual search. In *ICIP*, 2010.

[9] V. Chandrasekhar et al. Compressed histogram of gradients: A low-bitrate descriptor. *Int. J. Comput. Vision*, 96(3):384–399, 2012.

[10] D. Chen et al. Residual enhanced visual vectors for on-device image matching. In *Asilomar Conference on Signals, Systems, and Computers*, 2011.

[11] L.-C. Dai et al. Imshare: instantly sharing your mobile landmark images by search-based reconstruction. In *ACM MM*, 2012.

[12] J. Deng et al. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[13] J. Deng et al. What does classifying more than 10,000 image categories tell us? In *ECCV*, 2010.

[14] M. Douze et al. Evaluation of gist descriptors for web-scale image search. In *CIVR*, 2009.

[15] M. Everingham et al. The pascal visual object classes challenge 2007 results, 2007.

[16] M. Everingham et al. The pascal visual object classes challenge 2010 results, 2010.

[17] R.-E. Fan et al. LIBLINEAR: A library for large linear classification. *JMLR*, 9, 2008.

[18] B. Girod et al. Mobile visual search. *Signal Processing Magazine, IEEE*, 28(4):61–76, 2011.

[19] G. Griffin et al. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[20] J.-F. He et al. Mobile product search with bag of hash bits and boundary reranking. In *CVPR*, 2012.

[21] H. Jégou et al. Aggregating local descriptors into a compact image representation. In *CVPR*, 2010.

[22] S. Lazebnik et al. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[23] J.-G. Li et al. Face recognition using feature of integral gabor-haar transformation. In *ICIP*, 2007.

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

[25] K. Mikolajczyk et al. A comparison of affine region detectors. *Int. J. Comput. Vision*, 65(1-2):43–72, 2005.

[26] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *Int. J. Comput. Vision*, 60(1):63–86, 2004.

[27] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3):145–175, 2001.

[28] F. Perronnin et al. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.

[29] J. Sánchez and F. Perronnin. High-dimensional signature compression for large-scale image classification. In *CVPR*, 2011.

[30] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *ICCV*, 2003.

[31] A. Torralba et al. 80 million tiny images: A large data set for nonparametric object and scene recognition. *PAMI*, 30(11):1958–1970, 2008.

[32] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms, 2008.

[33] J. Wang and S.-F. Chang. Sequential projection learning for hashing with compact codes. In *ICML*, 2010.

[34] J. Wang et al. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[35] Y. Weiss et al. Spectral hashing. In *NIPS*, 2008.

[36] X. Yang and K.-T. Cheng. Accelerating surf detector on mobile devices. In *ACM MM*, 2012.