What Makes an Image Popular?

Aditya Khosla
Massachusetts Institute of Technology
khosla@csail.mit.edu

Atish Das Sarma
eBay Research Labs
atish.dassarma@gmail.com

Raffay Hamid
DigitalGlobe
raffay@gmail.com

ABSTRACT

Hundreds of thousands of photographs are uploaded to the internet every minute through various social networking and photo sharing platforms. While some images get millions of views, others are completely ignored. Even from the same users, different photographs receive different number of views. This begs the question: What makes a photograph popular? Can we predict the number of views a photograph will receive even before it is uploaded? These are some of the questions we address in this work. We investigate two key components of an image that affect its popularity, namely the image content and social context. Using a dataset of about 2.3 million images from Flickr, we demonstrate that we can reliably predict the normalized view count of images with a rank correlation of 0.81 using both image content and social cues. In this paper, we show the importance of image cues such as color, gradients, deep learning features and the set of objects present, as well as the importance of various social cues such as number of friends or number of photos uploaded that lead to high or low popularity of images.

1. INTRODUCTION

Over the last decade, online social networks have exploded in terms of number of users, volume of activities, and forms of interaction. In recent years, significant effort has therefore been expended in understanding and predicting online behavior, surfacing important content, and identifying viral items. In this paper, we focus on the problem of predicting popularity of images.

Hundreds of thousands of photographs are uploaded to the internet every minute through various social networking and photo sharing platforms. While some images get millions of views, others are completely ignored. Even from the same users, different photographs receive different number of views. This begs the question: What makes a photograph popular? Can we predict the number of views a photograph will receive even before it is uploaded?

WWW'14, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2744-2/14/04.
http://dx.doi.org/10.1145/2566486.2567996.

We investigate two crucial attributes that may affect an image's popularity, namely the image content and social context. In the social context, there has been significant work in viral marketing strategies and influence propagation studies for online social networks [10, 46, 37, 9, 24]. However, most of these works adopt a view where the spread of a piece of content is primarily due to a user viewing the item and potentially sharing it. Such techniques adopt an algorithmic view and are geared towards strategies for maximizing influence. On the contrary, in this work, we use social signals such as number of friends of the photo's uploader, and focus on the *prediction* problem of overall popularity.

Previous works have focused primarily on predicting popularity of text [42, 20] or video [43, 49, 38] based items. Such research has also explored the social context as well as the content of the text itself. However, image content is significantly harder to extract, and correlate with social popularity. Text based techniques can build on the wealth of methods developed for categorizing, NLP, clustering, and sentiment analysis. Comparatively, understanding such cues from image or video content poses new challenges. While there has been some work in video popularity prediction [17, 16], these tend to naturally focus on the social cues, comment information, and associated tags. From this standpoint, our work is the first to suitably combine contextual information from the uploader's social cues, and the content-based features, for images.

In order to obtain the image content features, we apply various techniques from computer vision and machine learning. While there has been a significant push in the computer vision community towards detecting objects [15, 5], identifying contextual relationships [45, 52], or classifying scenes [40, 34], little work has been expended towards associating key image components with 'global spread' or popularity in an online social platform. This is perhaps the first work that leverages image cues such as color histograms, gradient histograms, texture and objects in an image for ascertaining their predictive power towards popularity. We demonstrate through extensive exploration the independent benefits of such image cues, as well as social cues, and highlight the insights that can be drawn from either. We further show these cues combine effectively towards an improved popularity prediction algorithm. Our experiments illustrate several benefits of these two types of features, depending on the data-type distributions.

Using a dataset of millions of images from Flickr, we demonstrate that we can reliably predict the normalized view count of images with a rank correlation of up to 0.81 us-

ing both image content and social cues. We consider tens of thousands of users and perform extensive evaluation based on prediction algorithms applied to three different settings: *one-per-user*, *user-mix*, *user-specific*. In each of these cases, we vary the number of users, and the number of images per user. In each of these cases, the relative importance of different attributes are presented and compared against several baselines. We identify insights from our method that open-up several directions for further exploration.

We briefly summarize the **main contributions** of our work in the following:

- We initiate a study of popularity prediction for images uploaded on social networks on a massive dataset from Flickr. Our work is one of the first to investigate high-level and low-level image features and combine them with the social context towards predicting popularity of photographs.

- Combing various features, we present an approach that obtains more than 0.8 rank correlation on predicting normalized popularity. We contrast our prediction technique that leverages social cues and image content features with simpler methods that leverage color spaces, intensity, and simple contextual metrics. Our techniques highlight the importance of low-level computer vision features and demonstrate the power of certain semantic features extracted using deep learning.

- We investigate the relative importance of individual features, and specifically contrast the power of social context with image content across three different dataset types - one where each user has only one image, another where each user has several thousand images, and a third where we attempt to get specific predictors for users separately. This segmentation highlights benefits derived from the different signals and draws insights into the contributions of popularity prediction in comparison to simpler baseline techniques.

- As an important contribution, this work opens the doors for several interesting directions to pursue image popularity prediction in general, and pose broad social questions around online behavior, popularity, and causality. For example, while our work attempts to disentangle social and content based features, and derives new insights into their predictive power, it also begs the question on their impacts influencing each other through self-selection. Our work sheds some light on such interesting relations between features and popularity, but also poses several questions.

Paper Overview. The rest of the paper is organized as follows. We begin by mentioning related work in Section 2. Section 3 describes our problem formulation by providing intuition for what image popularity means, and also described the details of the dataset used throughout this paper. and methodologies. We then delve into the details of the prediction techniques using image content based cues and social cues independently in Sections 4 and 5 respectively. Our main combined technique, analysis and detailed experimental evaluation are presented in Section 6. Finally, Section 7 concludes with a summary of our findings and a discussion of several possible directions for future work.

2. RELATED WORK

Popularity prediction in social media has recently received a lot of attention from the research community. While most of the work has focused on predicting popularity of text content, such as messages or tweets on Twitter [42, 20], and some recent works on video popularity [43, 49, 38], significantly less effort has been expended in prediction of image popularity. The challenge and opportunity for images comes from the fact that one may leverage both social cues (such as the user's context, influence etc. in the social media platform), as well as image-specific cues (such as the color spectrum, the aesthetics of the image, the quality of the contrast etc.). Text based popularity prediction has of course leveraged the social context, as well as the content of the text itself. However, image content can be significantly harder to extract, and correlate with popularity.

Recently, there has been an increasing interest in analyzing various semantic attributes of images. One such attribute is image memorability which has been shown to be an intrinsic image property [21] with different image regions contributing differently to an image's memorability [29, 28]. Similarly, image quality and aesthetics are other attributes that have been recently explored in substantial detail [6, 8, 3]. Recent work has also analyzed the more general attribute of image intrestingness [53], particularly focusing on its correlation with image memorability [19]. There have also been a variety of other works dealing with facial [33], scene [41] and object [14] attributes.

In social context, there has been significant interest in understanding behavioral aspects of users online and in social networks. There is a large body of work studying the correlation of activity among friends in online communities; see examples in [18, 47, 48]. Most are forms of diffusion research, built on the premise that user engagement is contagious. As such, a user is more likely to adopt new products or behaviors if their friends do so [1, 36]; and large cascades of behavior can be triggered by the actions of a few individuals [18, 47]. A number of theoretical models have been developed to model influence cascades [37, 9]. The seminal work of Kempe et al. [24] also consider two models of influence; there has also been a long line of work on viral marketing starting from [10, 46]. All these works focus on influence but do not necessarily predict popularity beforehand - they are generic techniques that do not focus on a specific domain such as images in our context.

It would be interesting to explore emotions elicited from images, and their causal influence on image popularity on social networks - some interesting studies in other domains include [23, 30].

In the context of social networks, there has been a great deal of focus on understanding the effects of friends on behavior. In a very interesting piece of work, Crandall et al. [4] consider the two documented phenomenon, social influence and selection, in conjunction. They [4] suggest that users have an increased tendency to conform to their friends' interests. Another related work [31] considers homophily in social networks: they conclude that users' preference to similar people is amplified over time, due to biased selection. An interesting open question in our context is whether users, over time, have an increasing tendency to share images that cater to their friends' preferences - and if yes, how this affects overall popularity. As such, our work does not delve into these social aspects. We focus primarily on the task

Figure 1: Sample images from our image popularity dataset. The popularity of the images is sorted from more popular (left) to less popular (right).

of popularity prediction. It would be interesting to further explore the contributions of homophily to virality of images.

3. WHAT IS IMAGE POPULARITY?

There are various ways to define the popularity of an image such as the number of 'likes' on Facebook, the number of 'pins' on Pinterest or the number of 'diggs' on Digg. It is difficult to precisely pick any single one as the true notion of popularity - different factors are likely to impact these different measures of popularity in different contexts. In this work, we focus on the *number of views* on Flickr as our medium for exploring image popularity. Given the availability of a comprehensive API that provides a host of information about each image and user, together with a well established social network with significant public content, we are able to conduct a relatively large-scale study with millions of images and hundreds of thousands of users.

Figure 2(a) shows the histogram of the number of views received by the 2.3 million(M) images used in our study. Our dataset not only contains images that have received millions of views but also plenty of images that receive zero views. To deal with the large variation in the number of views of different images, we apply the log function as shown in Figure 2(b). Furthermore, as shown in [51], we know that unlike Digg, visual media tends to receive views over some period of time. To normalize for this effect, we divide the number of views by the duration since the upload date of the given image (obtained using Flickr API). The results are shown in Figure 2(c). We find that this resembles a Gaussian distribution of the view counts as one would expect. Throughout the rest of this paper, image popularity refers to this log-normalized view count of images.

In the following, we provide details regarding the datasets used (Section 3.1), and the evaluation metric (Section 3.2) for predicting image popularity.

3.1 Datasets

Figure 1 shows a sample of the images in our dataset. In order to explore different data distributions that occur naturally in various applications and social networks, we evaluate

[1]Note that the maximum view count of a single image in our dataset is 2.5M, but we truncate the graph on the left to amplify the remaining signal. Despite this, it is difficult to see any clear signal as most images have very few views. This graph is best seen on the screen with zoom.

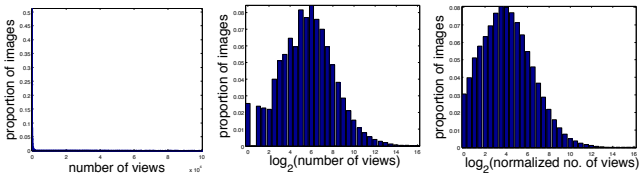

Figure 2: Histogram of view counts of images. The different graphs show different transformations of the data: (left) absolute view counts[1], (middle) $\log_2$ **of view counts** $+1$ **and (right)** $\log_2$ **view counts** $+1$ **normalized by upload date.**

our algorithms in 3 different settings namely *one-per-user*, *user-mix*, and *user-specific*. The main components that we vary across these settings is the number of images per user in the dataset, and whether we perform user-specific predictions. These settings are described below. In the later sections, we will show the importance of splitting image popularity into these different settings by illustrating how the relative contribution of both content-based and social features changes across these tasks.

One-per-user: For this setting, we use the Visual Sentiment Ontology dataset [3] consisting of approximately 930k images from about 400k users, resulting in a little over two images from each user on average. This dataset was collected by searching Flickr for 3244 adjective-noun-pairs such as 'happy guy', 'broken fence', 'scary cat', etc corresponding to various image emotions. This dataset represents the setting where different images belong to different users. This is often the case in search results.

User-mix: For this setting, we randomly selected about 100 users from the one-per-user dataset that had between 10k and 20k public photos shared on Flickr, resulting in a dataset of approximately 1.4M images. In this setting, we put all these images from various users together and perform popularity prediction on the full set. This setting often occurs on newsfeeds where people see multiple images from their own contacts or friends.

User-specific: For this setting, we split the dataset from the user-mix setting into 100 different users and perform training and evaluation independently for each user and average the results. Thus, we build user-specific models to predict the popularity of different images in their own collections. This setting occurs when users are taking pictures

or selecting pictures to highlight - they want to pick the images that are most likely to receive a high number of views.

3.2 Evaluation

For each of the settings described above, we split the data randomly into two halves, one for training and the other testing. We average the performance over 10 random splits to ensure the consistency of our results; overall, we find that our results are highly consistent with low standard deviations across splits. We report performance in terms of Spearman's rank correlation (ρ) between the predicted popularity and the actual popularity. Note that we use log-normalized view count of images as described in the beginning of this section for both training and testing. Additionally, we found that rank correlation and standard correlation give very similar results.

4. PREDICTING POPULARITY USING IMAGE CONTENT

In this section, we investigate the use of various features based on image content that could be used to explain the popularity of images. First, in Section 4.1 we investigate some simple human-interpretable features such as color and intensity variance. Then, in Section 4.2, we explore some low-level computer vision features inspired by how humans perceive images such as gradient, texture or color patches. Last, in Section 4.3, we explore some high-level image features such as the presence of various objects. Experimental results show that low-level computer vision features and high-level semantic features tend to be significantly more predictive of image popularity than simple image features.

4.1 Color and simple image features

Are simple image features enough to determine whether or not an image will be popular? To address this, we evaluate the rank correlation between popularity and basic pixel statistics such as the mean value of different color channels in HSV space, and intensity mean, variance, and skewness. The results are shown in Figure 3. We find that most simple features have very little correlation with popularity, with mean saturation having the largest absolute value of 0.05. Here, we use all 2.3M images independent of the settings described in Section 3.1. Since significant correlation does not exist between simple image features and popularity, we omit results from the individual settings for brevity.

Additionally, we look at another simple feature: the color histogram of images. As the space of all colors is very large ($256 * 256 * 256 \approx 16.8$m), and since small variations in lighting and shadows can drastically change pixel values, we group the color space into 50 distinct colors as described in [26] to be more robust to these variations. Then, we assign each pixel of the image to one of these 50 colors and form a ℓ_1-normalized histogram of colors. Using support vector regression (SVR) [12] with a linear kernel (implemented using LIBLINEAR [13]), we learn the importance of these colors in predicting image popularity[2]. The results are shown in Table 1 (column: *color histogram*), and visualized in Figure 4. Despite its simplicity, we obtain a rank correlation of 0.12 to 0.23 when using this feature on the three different data settings. We observe that on average, the greenish and bluish colors tend to have lower importance as compared to more reddish colors. This might occur because images containing more striking colors tend to catch the eye of the observer leading to a higher number of views.

[2]We find the hyperparameter $C \in \{0.01, 0.1, 1, 10, 100\}$ using five-fold cross-validation on the training set.

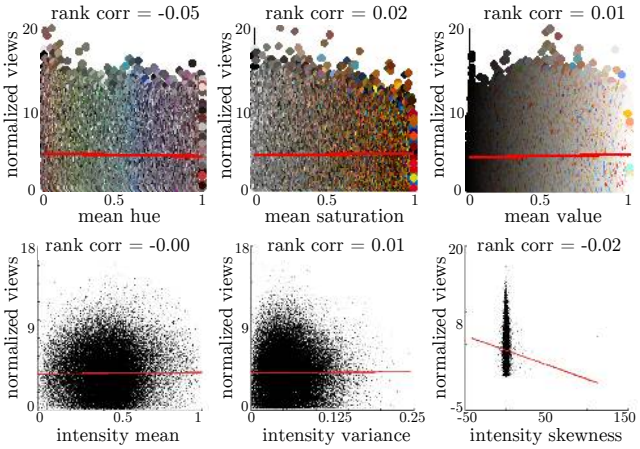

Figure 3: Correlation of popularity with different components of the HSV color space (top), and intensity statistics (bottom).

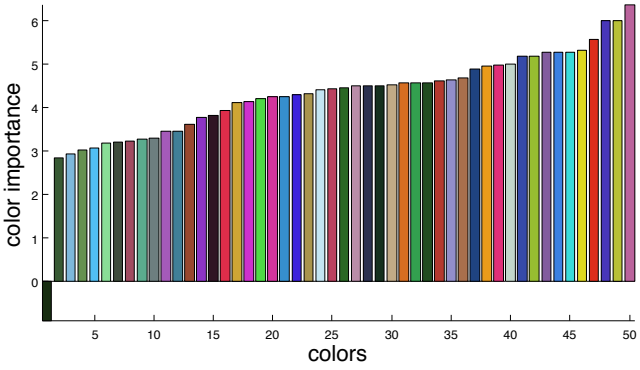

Figure 4: Importance of different colors to predict image popularity. The length of each bar shows importance of the color shown on the bar.

While the simple features presented above are informative, more descriptive features are likely necessary to better represent the image in order to make better predictions. We explore these in the remaining part of this section.

4.2 Low-level computer vision features

Motivated by Khosla et al. [29, 27], we use various low-level computer vision features that are likely used by humans for visual processing. In this work, we consider five such features namely gist, texture, color, gradient and deep learning features. For each of the features, we describe our motivation and the method used for extraction below.

Gist: Various experimental studies [44, 2] have suggested that the recognition of scenes is initiated from the encoding of the global configuration, or spatial envelope of the scene, overlooking all of the objects and details in the process. Essentially, humans can recognize scenes just by looking at

Dataset	Gist	Color histogram	Texture	Color patches	Gradient	Deep learning	Objects	Combined
One-per-user	0.07	0.12	0.20	0.23	0.26	0.28	0.23	**0.31**
User-mix	0.13	0.15	0.22	0.29	0.32	0.33	0.30	**0.36**
User-specific	0.16	0.23	0.32	0.36	0.34	0.26	0.33	**0.40**

Table 1: Prediction results using image content only as described in Section 4.

their 'gist'. To encode this, we use the popular GIST [40] descriptor with a feature dimension of 512.

Texture: We routinely interact with various textures and materials in our surroundings both visually, and through touch. To test the importance of this type of feature in predicting image popularity, we use the popular Local Binary Pattern (LBP) [39] feature. We use non-uniform LBP pooled in a 2-level spatial pyramid [34] resulting in a feature of 1239 dimensions.

Color patches: Colors are a very important component of human visual system for determining properties of objects, understanding scenes, etc. The space of colors tends to have large variations by changes in illumination, shadows, etc, and these variations make the task of robust color identification difficult. While difficult to work with, various works have been devoted to developing robust color descriptors [54, 26], which have been proven to be valuable in computer vision for various tasks including image classification [25]. In this paper, we use the 50 colors proposed by [26] in a bag-of-words representation. We densely sample them in a grid with a spacing of 6 pixels, at multiple patch sizes (6, 10 and 16). Then we learn a dictionary of size 200 and apply LLC [55] together with max-pooling in a 2-level spatial pyramid [34] to obtain a final feature vector of 4200 dimensions.

Gradient: In the human visual system, much evidence suggests that retinal ganglion cells and cells in the visual cortex V1 are essentially gradient-based features. Furthermore, gradient based features have been successfully applied to various applications in computer vision [5, 15]. In this work, we use the powerful Histogram of Oriented Gradient (HOG) [5] features combined with a bag-of-words representation for popularity prediction. We sample them in a dense grid with a spacing of 4 pixels for adjacent descriptors. Then we learn a dictionary of size 256, and apply Locality-Constrained Linear Coding (LLC) [55] to assign the descriptors to the dictionary. We finally concatenate descriptors from multiple image regions (max-pooling + 2-level spatial pyramid) as described in [34] to obtain a final feature of $10,752$ dimensions.

Deep learning: Deep learning algorithms such as convolutional neural networks (CNNs) [35] have recently become popular as methods for learning image representations [32]. CNNs are inspired by biological processes as a method to model the neurons in the brain, and have proven to generate effective representation of images. In this paper, we use the recently popular 'ImageNet network' [32] trained on 1.3 million images from the ImageNet [7] challenge 2012. Specifically, we use Decaf [11] to extract features from the layer just before the final 1000 class classification layer, resulting in a feature of 4096 dimensions.

Results: As described in Section 4.1, we train a linear SVR to perform popularity prediction on the different datasets. The results averaged over 10 random splits are summarized in Table 1. As can be seen from any of the columns, the rank correlation performance is best for the user-specific dataset, and next for user-mix, followed by one-per-user. However, it is important to note that the prediction accuracy when all the features are combined (as seen in the last column in Table 1) is at least 0.30 in all three cases. Given that these results only leverage image-content features (therefore ignore any aspects of the social platform or user-specific attributes), the correlation is very significant. While most of the current research focuses solely on the social network aspect when predicting popularity, this result suggests that the content also plays a crucial role, and may provide complementary information to the social cues.

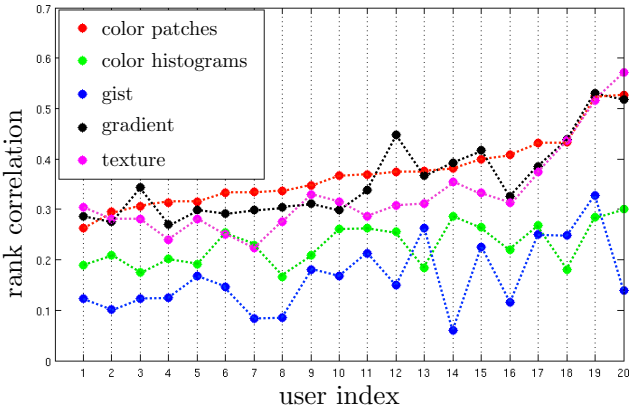

Figure 5: Prediction performance of different features for 20 random users from the *user-specific* setting. This figure is best viewed in color.

On exploring the columns individually in Table 1, we notice that the color histogram alone gives a fairly low rank correlation (ranging between 0.12 and 0.23 across the three datasets), but texture, and gradient features perform significantly better (improving the performance ranges to 0.20 to 0.32 and 0.26 to 0.34 respectively). The deep learning features outperform other features for the *one-per-user* and *user-mix* settings but not the *user-specific* setting. We then combine features by training a SVR on the output of the SVR trained on the individual features. This finally pushes the range of the rank correlations to 0.31 to 0.40.

In conjunction with the aggregate evaluation metrics described above and presented in Table 1, it is informative to look at the specific performance across 20 randomly selected users from the *user-specific* setting for each of these features - this is presented in a rank correlation vs. user index scatter plot in Figure 5. The performance for all features combined is not displayed here to demonstrate the specific variance across the image cues. As can be seen here, the color patches (red dots) performs best nearly consistently across all the users. For a couple of users, the gradient (black dots) performs better, and for others performs nearly as well as the color patches. In both these features, the prediction

Dataset	Mean Views	Photo Count	Contacts	Groups	Group members	Member duration	Is pro?	Tags	Title length	Desc. length	All
One-per-user	0.75	0.07	0.46	0.27	0.27	-0.08	0.22	0.52	0.21	0.45	**0.77**
User-mix	0.62	-0.05	0.26	0.24	0.06	-0.01	0.07	0.40	0.29	0.35	**0.66**
User-specific	n/a	n/a	n/a	n/a	n/a	n/a	n/a	0.19	0.16	0.19	**0.21**

Table 2: Prediction results using social content only as described in Section 5.

accuracy is fairly consistent. On the other hand, gist has a larger variance and the lowest average rank correlation. From the figure, it can be seen that texture (pink dots) also plays a vital role in predicting rank correlation accuracy, as it spans the performance range of roughly 0.22 to 0.58 across the randomly sampled users.

Another point of Figure 5 is to show that for personalization (i.e. user specific experiments), different features may be indicative of what a user's social network likes (because that is essentially what the model will capture by learning a user-specific model based on the current set of images). It is therefore interesting to note some variation in importance or relative ranking of features across the sampled users.

4.3 High-level features: objects in images

In this section, we explore some high-level or semantically meaningful features of images, namely the objects in an image. We want to evaluate whether the presence or absence of certain objects affects popularity e.g. having people in an image might make it more popular as compared to having a bottle. Given the scale of the problem, it is difficult and costly to annotate the dataset manually. Instead, we can use a computer vision system to roughly estimate the set of objects present in an image. Specifically, we use the deep learning classifier [32] described in the previous subsection that distinguishes between 1000 object categories ranging from lipstick to dogs to castle. In addition, this method achieved state-of-the-art classification results in the ImageNet classification challenge demonstrating its effectiveness in this task. We treat the output of this 1000 object classifier as features, and train a SVR to predict log-normalized image popularity as done in the previous sections. The results are summarized in Table 1.

We observe that the presence or absence of objects is a fairly effective feature for predicting popularity for all three settings with the best performance achieved in *user-specific* case. Further, we investigate the type of objects leading to image popularity. On the *one-per-user* dataset, we find that some of the most common objects present in images are: seashore, lakeside, sandbar, valley, volcano. In order to evaluate the correlation of objects with popularity, we compute the mean of the SVR weights across the 10 train/test splits of the data, and sort them. The resulting set of objects with different impact on popularity is as follows:

- **Strong positive impact:** miniskirt, maillot, bikini, cup, brassiere, perfume, revolver
- **Medium positive impact:** cheetah, giant panda, basketball, llama, plow, ladybug
- **Low positive impact:** wild boar, solar dish, horse cart, guacamole, catamaran
- **Negative impact:** spatula, plunger, laptop, golfcart, space heater

It is interesting to observe that this is similar to what we might expect. It is important to note that this object classifier is not perfect, and may often wrongly classify images to contain certain objects that they do not. Furthermore, there may be certain object categories that are present in images but not in the 1000 objects the classifier recognizes. This object-popularity correlation might therefore not pick up on some important object factors. However in general, our analysis is still informative and intuitive about what type of objects might play a role in a picture's popularity

5. PREDICTING POPULARITY USING SOCIAL CUES

While image content is useful to predict image popularity to some extent, social cues play a significant role in the number of views an image will receive. A person with a larger number of contacts would naturally be expected to receive a higher number of average views. Similarly, we would expect that an image with more tags shows up in search results more often (assuming each tag is equally likely). Here, we attempt to quantify the extent to which the different social cues impact the popularity of an image.

For this purpose, we consider several user-specific or social context specific features. We refer to user features as ones that are shared by all images of a single user. The user features that we use in our analysis are listed and described below. Note that this work investigates the relative merits of social and image features, and the goal isn't to heavily exploit one. Thus, we use relatively simple features for social cues that could likely be improved by using more sophisticated approaches.

- **Mean Views**: mean of number of normalized views of all public images of the given user
- **Photo count**: number of public images uploaded by the given user
- **Contacts**: number of contacts of the given user
- **Groups**: number of groups the given user belongs to
- **Group members**: average number of members in the groups a given user belongs to
- **Member duration**: the amount of time since the given user joined Flickr
- **Is pro**: whether the given user has a Pro Flickr account or not

We further subdivide some of the above features such as 'groups' into number of groups a given user is an administrator of, and the number of groups that are 'invite only'. We also include some image specific features that refer to the context (i.e. supporting information associated with the image as entered by the user but not its pixel content, as explored in Section 4). These are listed below.

- **Tags**: number of tags of the image
- **Title length**: length of the image title
- **Desc. length**: length of the image description

For each of the above features, we find its rank correlation with log-normalized image popularity. The results are shown in Table 2. Note that the user features such as Mean Views and Contacts would have the same value for all images by the particular user in the dataset. Not surprisingly, in the *one-per-user* dataset, the Mean Views feature performs extremely well in predicting the popularity of a new image with a rank correlation of 0.75. However, the rank correlation drops to 0.62 on the *user-mix* setting because there is no differentiation between a user's photos.

That said, 0.75 rank correlation is still a significantly better performance than we had expected since we note that the mean of views was taken over *all* public photos of the given user, not just the ones in our dataset. The mean number of public photos per user is over 1000, and of these, typically 1-2 are in our dataset, so this is a fairly interesting observation.

Another noteworthy feature is contacts - we see a rank correlation for these two datasets to be 0.46 and 0.26 respectively i.e. the more contacts a user has, the higher the popularity of their images. This is to be expected as their photos would tend to be highlighted for a larger number of people (i.e. their social network). Further, we observe that the image-specific social features such as tags, title length, and description length are also good predictors for popularity. As we can see in Table 2, their performance across the three dataset types range from 0.19 to 0.52 for tags, and 0.19 to 0.45 for description length. This is again to be expected as having more tags or a longer description/title increases the likelihood of these images appearing in the search results.

Further, we combine all the social features by training a SVR with all the social features as input. The results are shown in the rightmost column of Table 2. In this case, we see a rank correlation of 0.21, 0.66, and 0.77 for the *user-specific*, *user-mix*, and *one-per-user* datasets respectively. Thus, it is helpful to combine the social features, but we observe that the performance does not increase very significantly as compared to the most dominant feature. This suggests that many of these features are highly correlated and do not provide complementary information.

To contrast the results of social features from Table 2 with the image content features presented in the previous section in Table 1, we observe that the social features tend to perform better in the *one-per-user* and *user-mix* dataset types, while the image content features perform better in the *user-specific* dataset type. We suspect that in the user-specific dataset, where each user has thousands of images, the importance of personalized social cues becomes less relevant (perhaps due to the widespread range of images uploaded by them) and so the image content features become particularly relevant.

One thing that is evident from these results is that the social features and image content features are both necessary, and offer individual insights that are not subsumed by each other i.e. the choice of features for different applications largely depends on the data distribution. In the following section, we investigate techniques combining both of these features that turn out to be more powerful and perform well across the spectrum of datasets.

Dataset	Content only	Social only	Content + Social
One-per-user	0.31	0.77	**0.81**
User-mix	0.36	0.66	**0.72**
User-specific	0.40	0.21	**0.48**

Table 3: Prediction results using image content and social cues as described in Section 6.1.

6. ANALYSIS

In this section, we further analyze some of our results from the previous sections. First, we combine the signal provided by image content and social cues in Section 6.1. We observe that both of these modalities provide some complementary signal and can be used together to improve popularity prediction further. In Section 6.2 we visualize the good and bad predictions made by our regressors in an attempt to better understand the underlying model. Last, in Section 6.3 we show some preliminary visualizations of the image regions that make images popular by reversing the learned weights and applying them to image regions.

6.1 Combining image content and social cues

We combine the output of the image content and social cues using a SVR trained on the outputs of the most basic features, commonly referred to as late fusion. Table 3 shows the resulting performance. We observe that the performance improves significantly for all 3 datasets as compared to using either sets of features independently. The smallest increase of 0.04 rank correlation is observed for the *one-per-user* dataset as it already has a fairly high rank correlation largely contributed by the social features. This makes it difficult to improve performance further by using content features, but it is interesting to observe that we can predict the number of views in this setting with a fairly high rank correlation of 0.81. We observe the largest gains in the *user-specific* dataset, of 0.08 rank correlation, where image content features play a much bigger role as compared to social features.

6.2 Visualizing results

In Figure 6, we visualize some of the good and bad predictions made using our regressors. We show the four main quadrants: two with green background where the high or low popularity prediction matches the ground truth, and two with green background where the prediction is either too high or too low. We observe that images with low predicted scores (bottom half) tend to be less 'busy' and possibly lack interesting features. They tend to contain clean backgrounds with little to no salient foreground objects, as compared to the high popularity images. Further, we observe that the images with low popularity but predicted to have high popularity (top-left quadrant) tend to resemble the highly popular images (top-right quadrant) but may not be popular due to the social network effects of the user. In general, our method tends to do relatively well in picking images that could potentially have a high number of views regardless of social context.

6.3 Visualizing what makes an image popular

To better understand what makes an image popular, we attempt to attribute the popularity of an image to its re-

gions (similar to [29]). Being able to visualize the regions that make an image popular can be extremely useful in a variety of applications e.g. we could teach users to take better photographs by highlighting the important regions, or modify images automatically to make them more popular by replacing the regions with low impact on popularity.

In Figure 7, we visualize the contribution of different image regions to the popularity of an image by reversing the contribution of the learned weights to image descriptors. Since we use a bag-of-words descriptor, it can be difficult to identify exactly which descriptors in the image are contributing positively or negatively to the popularity score. Since max-pooling is used over spatial regions, we can carefully record the descriptors and their locations that led to the maximum value for each bag-of-words dictionary element. Then, we can combine this with the weights learned by SVR to generate a 'heatmap' of the regions that make an image popular. Note that this is a rather coarse heatmap because there can be image descriptors that have high values for certain dictionary elements, but not the highest, and their contribution is not considered in this max-pooling scenario. Thus, we end up with heatmaps that do not look semantically pleasing but indicate regions of high or low interest rather coarsely. This representation could be improved by using the recently popular mid-level features [50, 22] which encode more semantically meaningful structure.

From Figure 7, we can see that semantically meaningful objects such as people tend to contribute positively to the popularity of an image (first row right, and second row). Further we note that open scenes with little activity tend to be unpopular (with many exceptions of course). We observe that the number of high-scoring red/yellow regions decrease as the popularity of an image decreases (bottom row). Further, we observe that several semantically meaningful objects in the images are highlighted such as the train or different body parts, but due to the shortcoming described earlier, the regions are incoherent and broken up into several parts. Overall, popularity is a difficult metric to understand precisely based on image content alone because social cues have a large influence on the popularity of images.

Figure 7: Popularity score of different image regions for images at different levels of popularity: high (top row), medium (middle row) and low (bottom row). The importance of the regions decreases in the order red > green > blue.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we explored what makes images uploaded by users popular among online social media. Specifically, we explored millions of images on Flickr uploaded by tens of thousands of users and studied the problem of predicting the popularity of of the uploaded images. While some images get millions of views, others go completely unnoticed. This variation is noticed even among images uploaded by the same user, or images from the same genre. We designed an approach that leverages social cues as well as image content features to come up with a prediction technique for overall popularity. We also show key insights from our method that suggest crucial aspects of the image that determine or influence popularity. We extensively test our methodology across different dataset types that have a variable distribution of images per user, as well as explore prediction models that are focused on certain user groups or independent users. The results show interesting variation in importance of social cues such as number of photos uploaded or number of contacts, and contrast it with image cues such as color or gradients, depending on the dataset types.

Several directions remain for future exploration. An interesting question is predicting *shareability* as opposed to *popularity*. Are these different traits? There might be some images that are viewed/consumed, but not necessarily shared with friends. Does this perhaps have a connection with the emotion that is elicited? For example, peaceful images may get liked, funny images may get shared, and scary/disturbing images may get viewed but not publicly broadcasted. It would be interesting to understand the features/traits that distinguish the kinds of interaction they elicit from users. vs. those that elicit more uniform responses.

On a more open-ended note, do the influence of social context and image content spill across their boundaries? It is conceivable that a user who uploads refined photographs, over time, accumulates a larger number of followers. This could garner a stronger influence through the network and thereby result in increased popularity of photos uploaded by this user. Attribution might be inaccurate as a consequence - the resulting popularity may be ascribed to the user's social context and miss the image content. Popularity *prediction* as such may not be adversely affected, but what is the right causality here? Disentangling these features promises for an exciting direction to take this research forward. Being able to disentangle these factors may also result in improved content based popularity prediction by removing *noise* from the labels caused by social factors. Another specific question, for which data is unfortunately unavailable, is understanding time series of popularity for images: rather than simply looking at total popularity (normalized or unnormalized), can one investigate temporal gradients as well? For example, the total popularity of two images (or classes of images) may be the same, yet, one may have rapidly gained popularity and then sharply fallen, while another might have slowly and constantly retained popularity. These could perhaps exhibit fundamentally different photograph-types, perhaps the former being due to a sudden news or attention on an event, figure, or location, while the latter due to some intrinsic lasting value. Such exploration would be really valuable and exciting if the time series data were available for uploaded images.

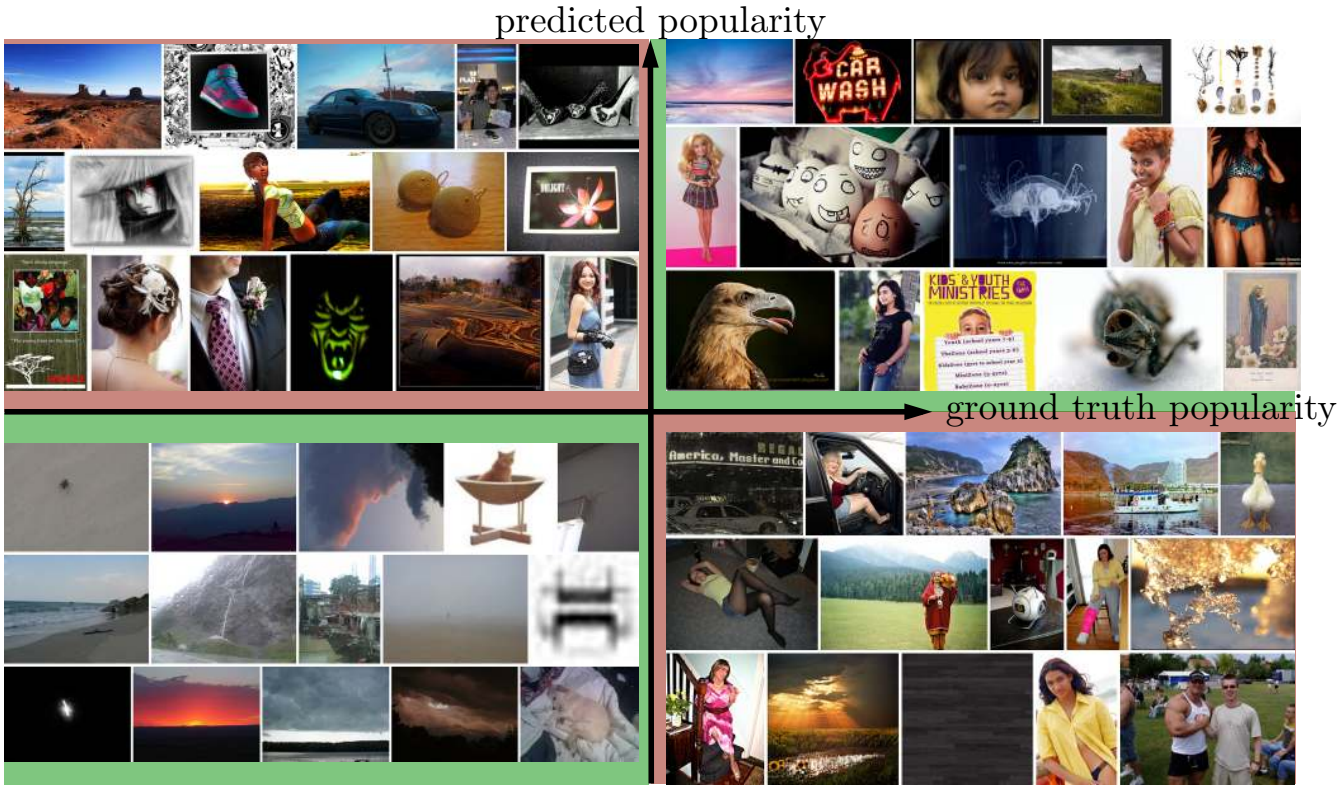

Figure 6: Predictions on some images from our dataset using Gradient based image predictor. We show four quadrants of ground truth popularity and predicted popularity. The green and red background colors represent correct and false predictions respectively.

Finally, from an application standpoint, is there a photography popularity tool that could be built here? Can photographers be aided with suggestions on how to modify their pictures for broad appeal vs artistic appeal? This could be an interesting research direction as well as a promising product. This is to be contrasted with some recent work[3] on aided movie script writing tools, where machine learning is potentially used to predict the likelihood of viewers enjoying the movie plot.

Acknowledgments

Aditya Khosla is supported by a Facebook Fellowship.

8. REFERENCES

[1] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD '06*, pages 44–54, 2006.

[2] I. Biederman. Aspects and extensions of a theory of human image understanding. *Computational processes in human vision: An interdisciplinary perspective*, pages 370–428, 1988.

[3] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*, 2013.

[4] D. J. Crandall, D. Cosley, D. P. Huttenlocher, J. M. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *KDD*, pages 160–168, 2008.

[3]http://www.nytimes.com/2013/05/06/business/media/solving-equation-of-a-hit-film-script-with-data.html

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[6] R. Datta, J. Li, and J. Z. Wang. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 105–108. IEEE, 2008.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[8] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1657–1664. IEEE, 2011.

[9] P. S. Dodds and D. J. Watts. A generalized model of social and biological contagion. *Journal of Theoretical Biology*, 2005.

[10] P. Domingos and M. Richardson. Mining the network value of customers. In *KDD*, pages 57–66, 2001.

[11] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[12] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. *Advances in neural information processing systems*, pages 155–161, 1997.

[13] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[14] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE, 2009.

[15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.

[16] F. Figueiredo. On the prediction of popularity of trends and hits for user generated videos. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 741–746. ACM, 2013.

[17] F. Figueiredo, F. Benevenuto, and J. M. Almeida. The tube over time: characterizing popularity growth of youtube videos. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 745–754. ACM, 2011.

[18] M. Gladwell. *The Tipping Point: How Little Things Can Make a Big Difference*. Back Bay Books, 2002.

[19] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Van Gool. The interestingness of images. In *IEEE ICCV*, 2013.

[20] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In *WWW (Companion Volume)*, pages 57–58, 2011.

[21] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 145–152. IEEE, 2011.

[22] M. Juneja, A. Vedaldi, C. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 2013.

[23] S. D. Kamvar and J. Harris. We feel fine and searching the emotional web. In *WSDM*, pages 117–126, 2011.

[24] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.

[25] F. S. Khan, J. Weijer, A. D. Bagdanov, and M. Vanrell. Portmanteau vocabularies for multi-cue image representation. In *Advances in neural information processing systems*, pages 1323–1331, 2011.

[26] R. Khan, J. Van de Weijer, F. S. Khan, D. Muselet, C. Ducottet, and C. Barat. Discriminative color descriptors. *CVPR*, 2013.

[27] A. Khosla, W. A. Bainbridge, A. Torralba, and A. Oliva. Modifying the memorability of face photographs. In *International Conference on Computer Vision (ICCV)*, 2013.

[28] A. Khosla, J. Xiao, P. Isola, A. Torralba, and A. Oliva. Image memorability and visual inception. In *SIGGRAPH Asia 2012 Technical Briefs*. ACM, 2012.

[29] A. Khosla, J. Xiao, A. Torralba, and A. Oliva. Memorability of image regions. In *Advances in Neural Information Processing Systems*, pages 305–313, 2012.

[30] S. Kim, J. Bak, and A. H. Oh. Do you feel what i feel? social aspects of emotions in twitter conversations. In *ICWSM*, 2012.

[31] G. Kossinets and D. J. Watts. Origins of homophily in an evolving social network1. *American Journal of Sociology*, 115(2):405–450, 2009.

[32] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1106–1114, 2012.

[33] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 365–372. IEEE, 2009.

[34] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006.

[35] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361, 1995.

[36] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1), 2007.

[37] M. E. J. Newman. Spread of epidemic disease on networks. *Physical Review E*, 2002.

[38] A. O. Nwana, S. Avestimehr, and T. Chen. A latent social approach to youtube popularity prediction. *CoRR*, abs/1308.1418, abs/1308.1418, 2013.

[39] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[40] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[41] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2751–2758. IEEE, 2012.

[42] S. Petrovic, M. Osborne, and V. Lavrenko. Rt to win! predicting message propagation in twitter. In *ICWSM*, 2011.

[43] H. Pinto, J. M. Almeida, and M. A. Gonçalves. Using early view patterns to predict the popularity of youtube videos. In *WSDM*, pages 365–374, 2013.

[44] M. Potter. Meaning in visual search. *Science*, 1975.

[45] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[46] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD*, pages 61–70, 2002.

[47] E. M. Rogers. *Diffusion of Innovations*. Simon and Schuster, 2003.

[48] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion. In *WWW '11*, 2011.

[49] D. A. Shamma, J. Yew, L. Kennedy, and E. F. Churchill. Viral actions: Predicting video view counts using synchronous sharing behaviors. In *ICWSM*, 2011.

[50] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *Computer Vision–ECCV 2012*, pages 73–86. Springer, 2012.

[51] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.

[52] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 273–280. IEEE, 2003.

[53] N. Turakhia and D. Parikh. Attribute dominance: What pops out? In *IEEE ICCV*, 2013.

[54] J. Van De Weijer, C. Schmid, and J. Verbeek. Learning color names from real-world images. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[55] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3360–3367. IEEE, 2010.