Multimedia Features for Click Prediction of New Ads in Display Advertising

Haibin Cheng, Roelof van Zwol, Javad Azimi[+], Eren Manavoglu
Ruofei Zhang, Yang Zhou, Vidhya Navalpakkam
Yahoo! Labs, Santa Clara, CA
Oregon State University, Corvallis, OR[+]
{hcheng,roelof,erenm,rzhang,yangzhou,nvidhya}@yahoo-inc.com
azimi@eecs.oregonstate.edu[+]

ABSTRACT

Non-guaranteed display advertising (NGD) is a multi-billion dollar business that has been growing rapidly in recent years. Advertisers in NGD sell a large portion of their ad campaigns using performance dependent pricing models such as cost-per-click (**CPC**) and cost-per-action (**CPA**). An accurate prediction of the probability that users click on ads is a crucial task in NGD advertising because this value is required to compute the expected revenue. State-of-the-art prediction algorithms rely heavily on historical information collected for advertisers, users and publishers. Click prediction of new ads in the system is a challenging task due to the lack of such historical data. The objective of this paper is to mitigate this problem by integrating multimedia features extracted from display ads into the click prediction models. Multimedia features can help us capture the attractiveness of the ads with similar contents or aesthetics. In this paper we evaluate the use of numerous multimedia features (in addition to commonly used user, advertiser and publisher features) for the purposes of improving click prediction in ads with no history. We provide analytical results generated over billions of samples and demonstrate that adding multimedia features can significantly improve the accuracy of click prediction for new ads, compared to a state-of-the-art baseline model.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

Keywords

Multimedia Features, Image, Flash, Click Prediction, New Ads, Display Advertising, GMM

1. INTRODUCTION

Display advertising generates revenue by showing graphical ads on web pages. It is traditionally sold as a guaranteed delivery (GD) contract, which constitutes a deal between a publisher and an advertiser to deliver a pre-specified number of impressions over a certain period of time at a fixed price per impression (CPM). An alternative mechanism for delivery that has been growing in the recent years is spot markets, where the impressions are sold one at a time. Spot markets, such as Right Media Exchange (Yahoo), lets the advertisers bid differently for each impression, allowing them to use highly granular targeting methods. Most spot markets also provide advertisers with a wider range of pricing models. Similar to GD, advertisers can choose to pay per impression (CPM). However many advertisers would prefer to pay only if the ad attracted the user's attention. To address this concern, many NGD exchanges provide performance based pricing models such as pay-per click (CPC) and pay-per-conversion (CPA), which can be further categorized as post-view or post-click, depending on there being a click before the conversion event. In a marketplace where ads with different payment models are competing for the same opportunity, the auction mechanism needs to convert the bids to a common currency. This is often done by computing expected revenue (eCPM). For a CPM ad the expected revenue is obviously going to be the same as the bid. For a CPC ad, however, the expected revenue depends on the probability that the user will click on that ad. Similarly, the expected revenue of a post-view CPA ad depends on the probability of conversion after the user views the ad; and for post-click CPA the expected revenue calculation needs to take into account both the click and the conversion probability. As a result, accurate prediction of click probability plays an important role in NGD advertising.

A state-of-the-art NGD system typically relies on machine learning models to estimate the click probability of eligible CPC/CPA ads. These models are trained using data collected from live systems. The identity of the users, publishers and ads are typically used as features in such models, together with some high level category information. For ads that have been in the system for a long period of time, the estimation of click probability using identifier based features is generally reliable, however it becomes a challenging problem for new ads. Besides, identifier based features for advertisers fail to provide any information about the aesthetics or the content of the ads, which is the key factor that the user responds to.

KDD '12 Beijing, China

This cold-start problem is often addressed by the use of content based features. However, most literature in this area focuses on textual representation of content. What makes this a unique problem for display advertising is the fact that the ads are not represented as text; display ads are graphical ads. In this paper we propose the use of multimedia features to represent the content (and aesthetics) of such ads. The contributions of our paper are as follows: We develop features that can be extracted from both static images as well as animated flash ads. We explore the use of global features to capture the visual characteristics of the entire image, local features to describe the sub sections in an image, and high level features to capture the user's perception. For flash ads, we develop an additional set of features that are extracted from the meta data. We also propose a clustering model based approach to capture the shared visual content of these images, and investigate the use of the group id as features. A feature selection algorithm is also developed to remove features with low click relevancy and high redundancy. Finally, we show experimental results on data sets collected from a commercial NGD exchange demonstrating multimedia features significantly improve the click prediction for new ads, compared to a state-of-the-art baseline model.

The paper is organized as follows. In Section 2, we describe the baseline NGD click prediction model. The multimedia features used in our work are introduced in Section 3 followed by the introduction of our feature selection method in Section 4. We present the feature analysis and modeling results in Section 5. Related work is discussed in Section 6.

2. NGD CLICK PREDICTION

In NGD advertising a spot auction is run for every ad slot on the publisher's page, in which all advertisers with matching target profiles participate. The ads are ranked based on their expected revenue and the winning ad is displayed. Estimating the expected revenue for pay-per click and post-click conversion payment models requires knowing the probability that the user will click on the candidate ad if shown in that ad slot on the publisher's page.

2.1 Model

The click prediction problem in NGD can be formulated as a classification problem, where each data point represents a publisher-ad pair presented to the user. Assume there is a set of n training samples, $\mathcal{D} = \{(\mathbf{f}(p_j, a_j, u_j), c_j)\}_{j=1}^{n}$, where $\mathbf{f}(p_j, a_j, u_j) \in \Re^d$ represents the d-dimensional feature space for publisher-ad-user tuple j and $c_j \in \{-1, +1\}$ is the corresponding class label ($+1$: click or -1 : no-click). Given a publisher p, ad a and user u, the problem is to calculate the probability of click $p(c|p, a, u)$. We use a maximum entropy algorithm for this supervised learning task because of its simplicity and strength in combining diverse features and large scale learning [3]. The maximum-entropy model, also known as logistic regression, has the following form:

$$p(c|p, a, u) = \frac{1}{1 + \exp(\sum_{i=1}^{d} w_i f_i(p, a, u))} \quad (1)$$

where $f_i(p, a, u)$ is the i-th feature derived from the publisher-ad-user tuple (p, a, u) and $w_i \in \mathbf{w}$ is the weight associated with it. Given the training set $\mathcal{D}$, the model learns the weight vector $\mathbf{w}$ by minimizing the total losses in the data formulated as:

$$LOSS(\mathbf{w}) = \sum_{i}^{n} L(\mathbf{w}; f_i(p_i, a_i, u_i), c_i) + \frac{\lambda}{2} \|\mathbf{w}\|^2 \quad (2)$$

where $L()$ is a logistic loss function used in this paper and λ controls the degree of $L2$ regularization to smooth the objective function. L-BFGS [5] is well suited to solve this kind of large scale convex optimization problem.

2.2 Features

Designing informative features is very important for supervised learning algorithms. Many features are derived from the publisher-ad-user tuple. On the user side, demographic information such as age and gender are common features used for click prediction. On the publisher and advertiser side there are hierarchies of entities. The entity identifiers are typically used as features in click models to capture the click behavior at different levels of abstraction. Publishers use site id to label their sites and section id to tag different parts of their pages. The *url* and *host* of the page are also informative features. An advertiser can set up multiple campaigns and creatives and the same creative can be used in multiple campaigns. And finally, publishers and advertisers connect to ad exchanges via networks, which constitute the root of the hierarchies.

The identifier features are all binary indicators that take the value 1 when present and 0 otherwise. Other ad features that are useful include the size of the ad, the topical category and the format (e.g. pop-up, floating or static banner ads).

Conjunctions are used to capture the interaction between different feature groups, such as user and publisher, publisher and ad, and user and ad conjunctions The number of features will grow exponentially after feature conjunction. Given a large set of identifiers, the final number of parameters in the model will be very large. Feature hashing [29] is a simple and effective dimension reduction technique to limit the feature space as well as maintaining the model performance by hashing the feature to a predefined number of bins.

3. MULTIMEDIA FEATURES

Display ads are mainly created in two formats, image and flash, from which we derive four sets of features. The first set is composed of features generated from images and image elements in flash ads. They are designed to capture the visual aspects of the images that may strongly affect users' response to the ads. The second set of features consists of meta information extracted from flash ads. In addition, we extract the latent mixture components from the images to capture their shared visual content as a separate set of features for click prediction. Finally, the image and flash features extracted from the ads are conjoined with user and publisher side features. For instance, the user age and ad color can be used as an additional feature to capture the variations in different age groups' responses to different colors.

3.1 Image Features

A digital image with resolution $X \times Y$ is typically treated as a grid of pixels with X rows and Y columns. The intensity of each pixel at location (x, y) can be represented in various color spaces including but not limited to RGB,

Grayscale, HSV, HSL and YUV. RGB stores individual values for red (R), green (G) and blue (B) for each pixel at (x, y). RGB can be easily converted to grayscale and consequently to binary, black or white, by setting a threshold value on grayscale value. HSV (hue, saturation, value) is another color space that takes human perception into account in the color encoding. In HSV, the brightness of a pure color is equal to the brightness of white. HSL (hue, saturation, lightness/luminance) is similar to HSV, except that the lightness of a pure color is equal to the lightness of a medium gray. The YUV model defines a color space in terms of one luma (Y) and two chrominance (UV) components. Different color spaces characterize an image from different perspective, based on which we can extract various features to describe the content of the image.

The features extracted from the image are divided into three categories, global features, local features and high level features. Global features are utilized to describe the content of the entire image using a small number of values while local features represent the characteristics of the local regions of the image. Both global and local features are computed directly from the image. The high-level features attempt to capture the human visual perception of the image. They involve more complex processing of the underling image data, that typically requires applying a model trained on an additional image corpus. A subset of these high-level visual features has been used in the context of predicting a user's photo preferences in social media sharing sites [28].

3.1.1 Global Features

Global features capture the visual effect of the entire image as a whole and are generally easy to calculate. The global features used in our work are introduced as following:

- **Brightness** of an image can be derived directly from two color spaces, the YUV color space [23] where "Y" stands for the luma component (the brightness) and HSL color space [18] where "L" measures the lightness of the image. The average, standard deviation, maximum and minimum of the luminance and lightness values of all the pixels in the image are derived as brightness features.

- **Saturation** measures the vividness of an image, whose value can be established directly from the HSV or HSL color space. Similar to brightness, we calculate the average, standard deviation, maximum and minimum of the saturation of all the pixels in the image.

- **Colorfulness** of an image is a measure of its difference against gray, which is calculated in RGB [12] as:

$$\begin{aligned} CF &= \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3\sqrt{\mu_{rg}^2 + \mu_{yb}^2} \\ rg &= R - G \\ yb &= \frac{R+G}{2} - B \end{aligned} \tag{3}$$

- **Naturalness** is the degree of correspondence between images and human perception of reality. Huang et al. [15] proposed an approach to obtain a quantitative description of Naturalness by grouping the pixels with $20 \leq L \leq 80$ and $S > 0.1$ in HSL color space according to their hue (H coordinate) value into three sets: Skin, Grass and Sky. The naturalness score $NS_{skin}, NS_{grass}, NS_{sky}$ as well as the proportional of pixels $NP_{skin}, NP_{grass}, NP_{sky}$ are derived from the image. The final naturalness score feature is calculated as:

$$NS = \sum_i NS_i \times NP_i, i \in \{\text{Skin, Grass, Sky}\} \tag{4}$$

- **Contrast** measures relative variation of luminance across the image in HSL color space. The simplest definition of contrast is the standard deviation of the luminance $L(x, y)$ of all image pixels. An extended version is to calculate the standard deviation of the normalized luminance $\frac{L(x,y)-L_{min}}{L_{max}-L_{min}}$ of all image pixels as the contrast.

- **Sharpness** measures the clarity level of detail of an image. Sharpness can be determined as a function of its Laplacian, normalized by the local average luminance in the surroundings of each pixel:

$$SH = \sum_{x,y} \frac{LP(x,y)}{\mu_{xy}}, LP(x,y) = \frac{\partial^2 L}{\partial x^2} + \frac{\partial^2 L}{\partial y^2} \tag{5}$$

  where μ_{xy} denotes the average luminance around pixel (x,y).

- **Texture** features correspond to human visual perception by capturing the spatial arrangement of color or intensities in an image. One of the most widely used texture feature set is developed by Tamura et al. [27], which achieved correspondences with psychological measurements successfully. Three out of the six Tamura textual features are utilized in our work, which describe the coarseness, contrast and directionality of the image.

- **Grayscale simplicity** features are extracted to represent the properties of the gray level image. Three features are extracted from the gray level histogram of the image consisting of 255 bins [2]. The first one calculates the contrast of the image by measuring the width of the gray level histogram which consists of 95% of the pixels in the image. The second feature counts the number of gray bins which contains the significant number of pixels. This feature measures the simplicity of the image in grayscale. The third feature simply calculates the standard deviation of the gray level values of all the pixels in the image. Javad et.al in [2] show that the proposed gray level features are very effective in predicting the CTR of ads.

- **RGB simplicity** features can represent the simplicity of a color image [22]. Similar to [2], we quantize the RGB space into 512 bins by dividing each channel into 8 equal intervals. The number of RGB bins whose number of pixels are above a certain threshold is calculated as the simplicity feature in RGB space. We also take out the RGB bin with the maximum number of pixels as the dominant color and calculate its ratio with regard to the total number of pixels in the image as another features. Two similar features can also be calculated in the HSV color space.

- **Color harmony** property of an image are assumed to be correlated with the appeal of an image to a random user [8]. Two features are extracted from the

image based on the 7 color harmonic distribution templates[16, 17] created from the hue value of HSV color space. From the HSV color space of an image, we calculate the average deviation from each color harmony template and the deviation from the best two fitted models are reported as two color harmony features [2].

- **Hue histogram** features are based on the hue value of all the pixels in image. Each Hue value in HSL or HSV color space represents a color by itself. Similar to [2], three features are extracted based on hue histogram of an image consisting of 20 bins. The first feature counts the number of bins including number of pixels more than a threshold value, indicated as number of significant hues. The second feature calculates the contrast of the hue histogram as the maximum arc length distance between any two significant bins. The third feature calculates the standard deviation of the hue arc length of all the pixels in the image, which shows the distribution of the hue color in the images.

3.1.2 Local Features

Users often pay more attention to certain regions in an image. Features that are generated from local regions can therefore be useful. To generate local features, we first divide the image into many segments using a connected component algorithm [25]. Let $S = S_1, S_2, ...S_g$ be the final g set of segments created from the image (note that segments smaller than 5% of the image are dropped). Some of the global features introduced in Section 3.1.1 can be extended to local features by calculating them in local segments similar to [2]. The local features used in our work are described as follows:

- **Basic segment statistics** features are extracted from the basic statistics of the segments. The first feature is simply the number of segments g in the image, which may indicate how busy an image is. Another feature is the contrast of segment sizes, which is calculated as the difference between the size of the largest and the smallest component. The third feature calculates the ratio of the largest connected color component to the whole image in terms of number of pixels. This feature will have a larger value for a smooth image. The fourth feature is defined as the rank of the hue bin, considering the bin size in descending order, associated with the largest connected component in the image. The last two features are calculated in the same way as the third and fourth feature except that they are based on the second largest connected component.

- **Segment hue histogram** is generated for each segment in the image. Similar to the global hue histogram features, six local hue features are extracted: 1) the number of significant hues in the image falling in the largest segment, 2) the number of significant hues in the largest segment, 3) the largest number of significant hues among all the segments, 4) the contrast of the number of significant hues among all segments, 5) the contrast of the hues in the largest segment, and 6) the standard deviation of all segment's hue contrasts.

- **Segment color harmony** features are similar to the global color harmony features except that they are computed on the largest segment. Two features are generated, the minimum deviation from the best fitted color harmony model and the average deviation of the best two fitted color harmony models.

- **Segment brightness** features are based on the lightness of each segments calculated in HSL color space. Three features are calculated, 1) the average lightness of all the pixels in the largest segment, 2)the standard deviation of average lightness among all the segments, and 3) the contrast of average lightness among all the segments.

3.1.3 High Level Features

Most of the global and local features introduced above describe the content of the image at low level of visual perception. In this section, we introduce a set of more advanced features that is able to capture high level perception or conception information of an image.

- **Interest points** are the pixels in the image that constitute the edges, e.g. high-contrast regions, of objects in an image. Lowe et al. [21] developed the SIFT algorithm (Scale-invariant feature transform) to identify the interesting points for object detection. We simply use the number of interesting points as a feature for our task, which may indicate the complexity of an image in terms of the number of objects it contains.

- **Saliency map** is a binary map detected from the image using saliency detection algorithms such as [14] to distinguish the objects from the background whose saliency value is less than a predefined threshold. Similar to [2], based on the saliency map, many features are extracted such as: 1) the ratio of background to the whole image, 2) the number of connected components of background, 3)the ratio of the largest connected component of background to the whole image, 4) the number of connect components in the saliency map, 5) the ratio of the largest connected saliency area to the whole image, 6) the average weight of the largest connected components of saliency map, 7) the distortion of the connected saliency areas calculated as the overall distance among all the components centroids, and 8) overall distance of all component centroids from the center of image.

- **Text** in an image can be extracted using standard OCR (Optical Character Recognition) algorithms [26]. Simple features such as the number of characters and number of words can be easily created afterward as two possible features. Note that, these two features are independent from the content of the text and therefore it does not need a very accurate OCR toolbox.

- **Human faces** in an image can be extracted by face detection algorithms [4]. For our ad click prediction task, we created four features using the recognized faces in the image: the number of profile faces, the proportion of profile faces in terms of pixels, the number of frontal faces, the proportion of frontal faces in terms of pixels.

3.2 Flash Features

Flash is commonly used in display advertising, because it enriches the user experience by supporting streamlining of video, audio, animation of text, drawings and images.

Compared with static image ads, flash ads are more attractive and therefore more likely to be clicked by users. Hence features extracted from flash ads can provide additional information to click models. In our work, a flash ad is decomposed into many elements including image, sound, font, text, button, shape, frame, and action. Obviously the image features introduced in Section 3.1 can be extracted from the image element of a flash ad. A list of flash specific features are introduced as following:

- **Basic flash** features are derived by counting the number of movie clips, shapes, fonts and frames in the flash.
- **Audio** feature is a binary feature indicating whether a flash ad contains audio. A flash ad with audio may be more attractive to users than soundless ads.
- **Text** features such as number of characters, number of words and a number of pre-determined keywords (e.g., "click", "free") are derived from the text elements in the flash.

3.3 Mixture Component Features

When users look at a display ad they do not perceive it as a matrix of pixels, but rather they process the content of the ad. Images with similar content may receive similar responses from users. One way to capture this is to cluster images based on content similarity and use the cluster membership as a feature. We use a Gaussian Mixture Component (GMM) model for this purpose, instead of more advanced models such as Probabilistic Latent Semantic Analysis (PLSA) [13], because of its scalability. The weight of mixtures as well as the mean vectors and covariance matrices for each mixture are learned through a maximum likelihood process from the images in the training set, which are described as vectors of image features introduced in Section 3.1. For every image in the test data we estimate the probability of component membership from the learned model. The component id with the maximum posterior probability is then used as the mixture component feature in the click model.

3.4 User and Publisher Conjunction Features

The "attractiveness" of a rich media ad to different users will vary since users may have different interests and taste. For example, male and female users will certainly react differently when seeing an ad with a beautiful human face; young users may be more attracted to ads with cartoons than older users. The ad performance on different publishers also varies. As an example, ads with cars in the image are more likely to be clicked when shown on a automobile related site than when shown on a fashion site. These factors are not taken into account by the multimedia features introduced above since they are extracted from the ad content only. Conjunction features solve this problem by taking the cross product of the user features (such as age, gender etc.) or publisher features (such as publisher id, url, etc.) with the multimedia features.

4. FEATURE SELECTION

Some of the multimedia features introduced above may not be relevant to the click prediction task or can be redundant because of the intrinsic correlations with other features. Including all of them in the model may not only increase the model complexity but also degrade the model performance. In this paper, a clustering based feature selection algorithm is developed to remove irrelevant and redundant features. First, features are ranked based on their relevance to the target in the training data, measured using a standard mutual information method [32]. Irrelevant features are removed by thresholding the relevance score.

We use a spectral clustering method to group similar features together. A fully connected similarity graph is constructed, in which nodes represent features and edge weights are defined by the similarity (in terms of mutual information) of the two features they connect. A normalized cut method [9] is then applied to the graph to obtain clusters of features that are strongly correlated with each other. Finally features within each cluster are ranked using the relevance score and only the top k features in each cluster are selected to build the model. Clearly, there is a trade-off between relevance and redundancy that can be tuned by varying the k according to the size of the cluster.

Table 1: Feature set in baseline model for click prediction.

Feature Group	Feature Name
Publisher	publisher id, publisher network id, section id, url, host
User	age,gender
Ad ID	advertiser id, campaign id, creative id, advertiser network id
Ad Type	ad size, offer type id, pop type id

5. EXPERIMENTS

In this section, we first describe the features, data sets and models used in the experimental evaluations. Then we present a click-through rate (CTR) analysis of the image features. Finally, we compare the click prediction model with multimedia features to a baseline model on click prediction of new ads.

5.1 Experimental Setup

We compare the models using the data sampled from the Yahoo! NGD advertising system for a period of 5 weeks. Each sample is an impression of an ad and is labeled as "clicked" or "not clicked", as derived from the user click logs. The data from first 4 weeks is used as training set and the last week is treated as test set. There are approximately 2.3 billion samples in the training data. The data contains about 1.4 million unique display ads. 54% of the ads are images and the rest 46% are flash objects. The first image of the flash ad is extracted as its image representation. In order to evaluate the performance of multimedia features on new ads, we further create subsets of test data, which include only impressions with new ads that never appeared in the training period.

The baseline model used for comparison is trained based on numerous features extracted from users, publishers and advertisers. Table 1 summarizes all the basic features used in the baseline model. The baseline model also includes all the conjunction features conjoining every pair of feature groups in Table 1. A 24-bit hash function is used to hash all the features into 16 million bins. With 16 million features, the baseline model is considered strong. To evaluate the

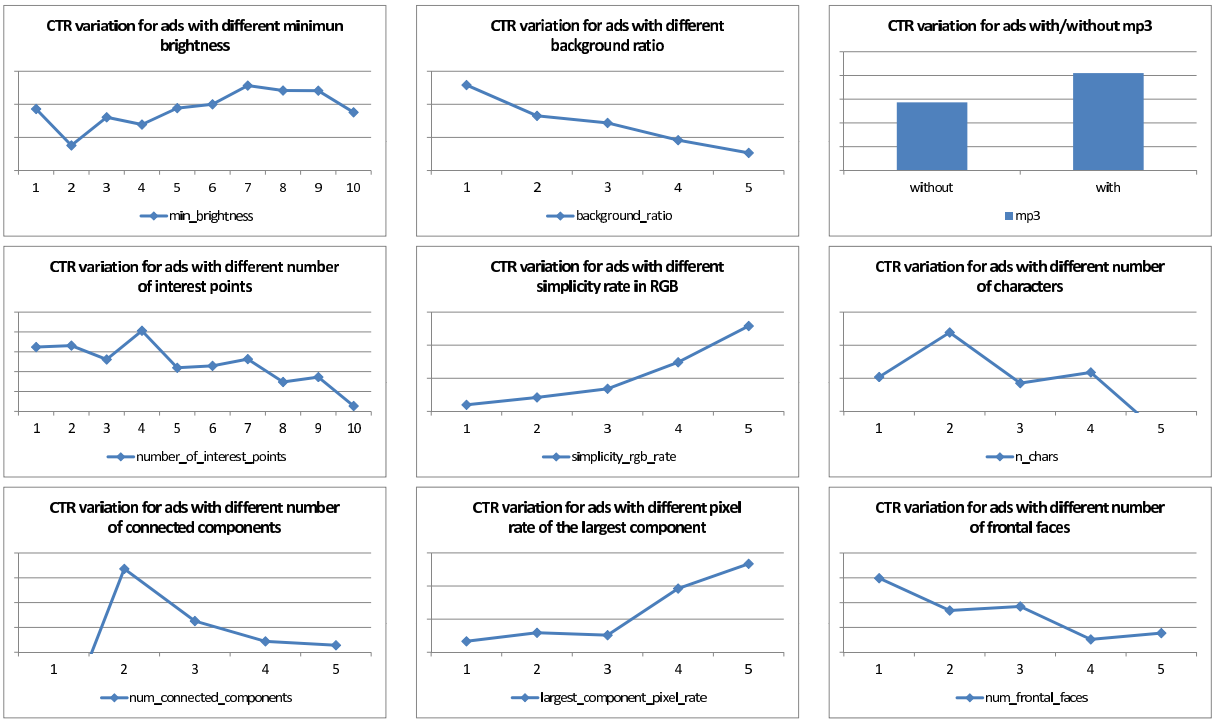

Figure 1: Variation of CTR with regard to different multimedia features

performance of multimedia features in click prediction, we retrain a model by adding multimedia features to each training sample as additional features. Events in the test set are ordered by the predicted clickability score. Precision-recall (**P-R**) curves as well as area under curve metric (**AUC**) are calculated to measure the accuracy of the models.

We evaluated three click prediction models with multimedia features. The first model uses all the multimedia features introduced in Section 3. The second model adds only the mixture component feature to the baseline model. The number of components is set to 150 in the experiment. The third model utilizes a subset of the multimedia features selected using the method presented in Section 4. Note that all the models use the same 24-bit hash function to hash all the features. As a result, adding multimedia features does not increase the total number of features.

5.2 Feature Analysis

Here we conduct an initial study on various multimedia features to obtain an intuitive understanding of their impact on ad click behavior. One week of test data is used for this analysis. We quantize each multimedia feature into multiple bins using a k-means clustering algorithm and the CTR for each bin is calculated. Figure 1 shows the CTR distribution for 9 representative multimedia features selected in our study. The Y-axes of each figure shows the CTR and the X-axes represents the bin index of features. The actual CTR values in the figures are omitted to protect proprietary information. Our observations from Figure 1 are listed as follows (note that images referred below include flash images):

- CTR increases nearly linearly with the minimum brightness of the ads.
- Large background image ads receive less clicks than small background image ads.
- Flash ads with audio generate more clicks than flash ads without audio.
- Ads with larger number of interest points have lower CTR than ads with a small number of interest points.
- When an image ad has more pixels of dominant color (simpler), it is likely to generate more clicks.
- There is a negative correlation between the CTR and the number of characters detected in the image.
- Image ads with a small number of connected components obtained from segmentation are generally preferred by users over image ads with a large number of connected components.
- Image ads whose largest connected component is big are likely to receive more clicks.
- There is a negative correlation between the CTR and the number of faces detected in the image.

It is clear that all these features are highly correlated with CTR and therefore have a great potential to improve the performance of click models. In fact, many of these observations are consistent with each other. For example, "simple" images often have a small number of interest points, a few connected components, or a high ratio of dominant color. These observations can also be used to guide the creative design process for the purpose of increasing their CTR.

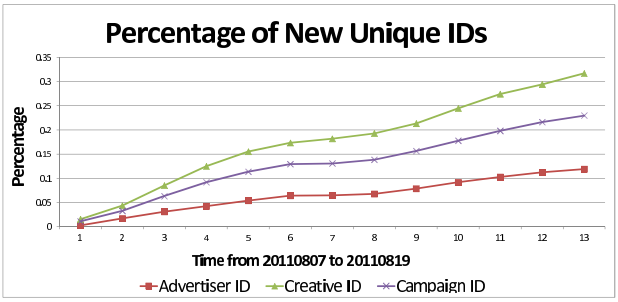

Figure 2: Percentage of new unique advertiser identifiers, creatives and campaigns emerging for each day in one month (relative to those existing in the training period).

5.3 Click Prediction for New Ads

Modeling multimedia features allows us to improve the click prediction for new ads by enabling the use of historical performance of pre-existing ads with similar content. In order to evaluate the scope of the problem we first measure the burst rate of new ads in the data. For this analysis we treat the first 4 weeks of data as the reference set, and calculate the ratio of new ads per day for the following 2 weeks. Advertiser campaigns often have an inherent structure: one advertiser can have multiple campaigns, and one campaign might have multiple creatives. In the case of Yahoo's exchange the same creative can actually appear in multiple campaigns. To understand the problem space better we measure the burst rate at each of these levels separately. As shown in Figure 2, it is clear that the burst rate is increasing steadily day by day for all three levels, however creative has the most significant change. 31.7% of unique ads are new creatives after a period of 13 days. The burst rate of campaigns is slower, with 23.0% of new unique ads after a period of 13 days. The advertiser is the most stable of the three levels, which still has 11.9% new unique ads. This underlines the impact of our work on the performance of the NGD advertising system.

Table 2: Performance improvement in terms of AUC of click prediction model with multimedia features on new ads with different level of newness.

Newness Level	AUC Gain
New Advertiser ID (newest)	6.50%
New Creative ID	2.15%
New Campaign ID	1.27%
New Campaign+Creative ID	3.31%

Then we evaluate the performance of the three click models with multimedia features as introduced in Section 5.1 on new ads. Specifically we create multiple slices of test data which contain new ads represented at four different levels: advertiser, creative, campaign and campaign×creative. The campaign×creative slice includes ads with either new campaigns or new creatives. The P-R curve and corresponding AUC for each slice of test data are shown in Figure 3. Looking at the test data with new advertiser id, we observe a significant improvement by adding the multimedia features. The model with all multimedia features gains 6.03% in terms of AUC. Using feature selection method introduced in Section 4 improves the performance even more, with a 6.50% gain over baseline. This is not surprising since a lot of features introduced in Section 3 seem strongly correlated with each other. It is also worth mentioning that the mixture component feature is so informative that adding it alone brings 1.5% gain over the baseline. We also observe notable improvements by adding multimedia features on the other new ad slices. The improvements are relatively smaller compared to the result on the new advertiser slice. This is expected because advertisers use the same advertiser identifier for new creatives and new campaigns thus providing some historical information even for new ads. Table 2 summarizes the AUC gains compared to the baseline for the model with selected multimedia features using the method introduced in Section 4. All the improvements are statistically significant ($p < 0.01$) based on a simple paired $t-$test on the prediction results generated by thresholding the prediction score of the compared models using the equal error point (EER). In general, multimedia features play a important role on advertisers with limited historical information. Note that all these improvements are achieved without adding model complexity in terms of the total number of features due to the feature hashing strategy.

6. RELATED WORK

Click prediction in on-line advertising has received increasing attention from the research community in recent years. Most of the work in the literature can be categorized broadly as either new feature or new model development. In terms of feature development, Liu et al. [20] propose to use syntactic features for sponsored search to model the relevance between query and ad by treating the ad's text as a short document and building a language model as in classic information retrieval; Chakrabarti et al. [6] developed click feedback features [6] based on aggregated historical click data, which has proven to be very effective in click prediction; Cheng et al. [7] developed a personalized click model by including user specific features and demographic features which dramatically improve the performance of click prediction. In terms of model development, maximum entropy [7, 3] or decision trees [10] are common models used for click prediction in on-line advertising. Specifically for sponsored search we see the use of generative graphical models [30] that focus on identifying the factors that affect the user's response to ads on a search page; Graepel et al. [11] propose a Bayesian on-line learning algorithm used for CTR prediction in Bing's sponsored search product.

Most published work in the area of click prediction for new ads is focused on identifying new ads either by using categories [6] or identifying a set of similar ads by using the textual content of the ads [10], which cannot be applied to display advertising directly. In our case, some level of ad and page category was already being utilized by the baseline models. An alternative approach was proposed by Agarwal et al. [1] that utilizes the hierarchy in the advertisers campaigns to improve the prediction for new ads. While we did not use the exact same model, we did use features that capture part of this information in our baseline models. Therefore the improvements we present in this paper are complementary to their work.

Image features have been widely researched for the task of content based image retrieval (CBIR)[31, 24] including both global features and local features. Global features charac-

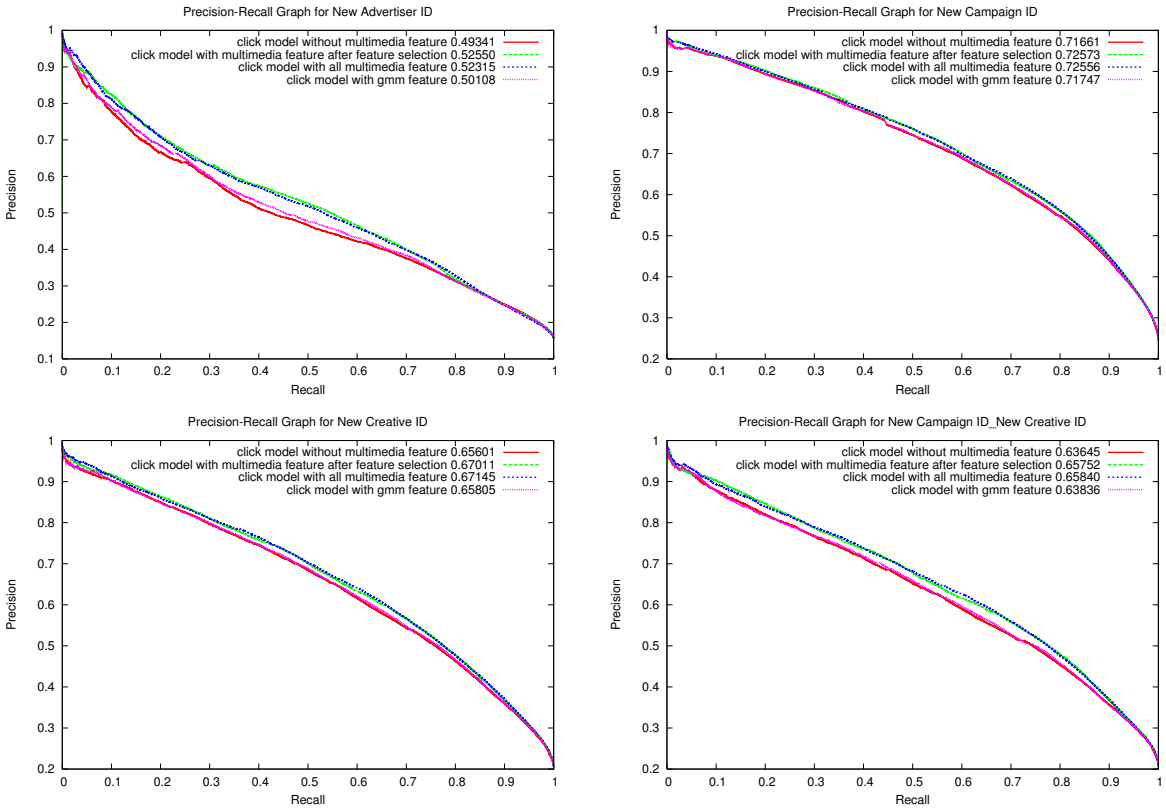

Figure 3: Variation of CTR with regard to different multimedia features

terize images using global characteristics such as color histogram, texture values, shape parameters. For example, Yoo et al. [31] developed a set of global visual features for the task of large scale image searching. Global features fail to work when the retrieval task is targeted to specific local objects, e.g. human face recognition. Local features are typically used [24] in such tasks. CBIR mainly focuses on finding similar images in a database, whereas our task requires measuring the "attractiveness" of the image particularly in terms of CTR. There is some work in literature developing features to estimate the beauty of an image using many features [33, 19], however little work is found on exploring their correlation with CTR. In addition to the features being studied in this paper, there are other derivatives from the computer vision domain that could have a strong impact on user's response to the ads, such as company logos. We would like to extend our analysis to include these features.

To the best of our knowledge Azimi et al.'s [2] proposal is the first use of multimedia features for the purposes of estimating click probability of display ads. Some of the image features described in Section 3.1 were introduced in their work. However their proposal was limited to static image ads only and the experimental results provided in their work compared to a simple weighted sampling based CTR estimation on a small set of samples.

7. CONCLUSIONS

In this paper, we propose to use multimedia features to improve the accuracy of click prediction for new ads in a NGD advertising system. Specifically, we developed image and flash features that describe the visual perception of the content of display ads. We analyzed the click distributions for the different multimedia features and observed a strong correlation between the CTR and features defined. We further tested the models with multimedia features on large scale data collected from real traffic. Compared to a state-of-the-art baseline model without multimedia features, we observed significant improvement in terms of prediction accuracy on test data with new ads. We also developed a feature selection algorithm to remove the redundant and highly correlated features, which was shown to improve the accuracy of the final prediction task.

8. ACKNOWLEDGMENTS

We thank our colleagues Kannan Achan, Lihong Li, Olivier Chapelle and Romer Rosales for their assistance with data collection and model discussions.

9. REFERENCES

[1] D. Agarwal, R. Agrawal, R. Khanna, and N. Kota. Estimating rates of rare events with multiple hierarchies through scalable log-linear models. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 213–222, New York, NY, USA, 2010. ACM.

[2] J. Azimi, R. Zhang, Y. Zhou, V. Navalpakkam, J. Mao, and X. Fern. The impact of visual appearance on user response in online display advertising. *CoRR*, 2012.

[3] A. L. Berger and V. J. D. Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22:39–71, 1996.

[4] G. Bradski and A. Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly Media, 2008.

[5] R. H. Byrd, J. Nocedal, R. B. Schnabel, R. H. B. J. Nocedal, and R. B. Representations of quasi-newton matrices and their use in limited memory methods. *Mathematical Programming*, 63:129–156, 1994.

[6] D. Chakrabarti, D. Agarwal, and V. Josifovski. Contextual advertising by combining relevance with click feedback. In *Proceeding of the 17th international conference on World Wide Web*, pages 417–426, New York, NY, USA, 2008. ACM.

[7] H. Cheng and E. Cantú-Paz. Personalized click prediction in sponsored search. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 351–360, New York, NY, USA, 2010. ACM.

[8] D. Cohen-Or, O. Sorkine, R. Gal, T. Leyvand, and Y.-Q. Xu. Color harmonization. In *ACM Transactions on Graphics*, pages 624–630.

[9] T. Cour, F. Benezit, and J. Shi. Spectral segmentation with multiscale graph decomposition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1124–1131, 2005.

[10] K. S. Dave and V. Varma. Learning the click-through rate for rare/new ads from similar ads. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 897–898, New York, NY, USA, 2010. ACM.

[11] T. Graepel, J. Q. Candel, T. Borchert, and R. Herbrich. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

[12] D. Hasler and S. E. Suesstrunk. Measuring colorfulness in natural images. *Human Vision and Electronic Imaging VIII, Proceedings of the SPIE*, 5007:87–95, 2003.

[13] E. Horster, R. Lienhart, and M. Slaney. Continuous visual vocabulary models for plsa-based scene recognition. In *ACM International Conference on Image and Video Retrieval*, pages 319–382, 2008.

[14] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *In IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[15] K.-Q. Huang, Q. Wang, and Z.-Y. Wu. Natural color image enhancement and evaluation algorithm based on human visual system. *Computer Vision and Image Understanding*, 103:52–63, 2006.

[16] J. Itten. The art of color. *New York: Van Nostrand Reinhold Company*, 1960.

[17] J. Itten. Color design. *Asakura Shoten*, 1995.

[18] G. H. Joblove and D. Greenberg. Color spaces for computer graphics. *Computer Graphics (SIGGRAPH '78 Proceedings)*, 12(3):20–25, Aug. 1978.

[19] N. Kalidindi, A. Le, J. Picone, H. Y. L. Zheng, and V. A. Rudis. Scenic beauty estimate of forestry images. In *Proceedings of the IEEE Southeastcon*, pages 337–339, 1997.

[20] C. Liu, H. Wang, S. Mcclean, J. Liu, and S. Wu. Syntactic information retrieval. In *Proceedings of the 2007 IEEE International Conference on Granular Computing*, page 703, Washington, DC, USA, 2007. IEEE Computer Society.

[21] D. G. Lowe. Object recognition from local scale-invariant features. *Proceedings of the International Conference on Computer Vision*, 2:1150–1157, 1999.

[22] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. *Proceedings of the 10th European Conference on Computer Vision: Part III*, pages 386–399, 2008.

[23] C. Poynton. Yuv and luminance considered harmful: A plea for precise terminology in video. *Engineering*, pages 1–4, 2008.

[24] Z. R. Srihari and A. Rao. Image background search: Combining object detection techniques with content based image retrieval(cbir) systems. In *Proceedings of the IEEE workshop on Content Based Access of Image and video Libraries*, pages 97–101, 1999.

[25] L. G. Shapiro and G. C. Stockman. *Computer Vision*. Prentice Hall, 2002.

[26] C. Y. Suen, M. Berthod, and S. Mori. Automatic recognition of handprinted characters — the state of the art. *Proceedings of the IEEE*, 68:469–487, 1980.

[27] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics*, 8:460–473, 1978.

[28] R. van Zwol, A. Rae, and L. G. Pueyo. Prediction of favourite photos using social, visual, and textual signals. In *ACM Multimedia*, pages 1015–1018, 2010.

[29] K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1113–1120, 2007.

[30] W. Xu, E. Manavoglu, and E. Cantú-Paz. Temporal click model for sponsored search. SIGIR '10, pages 106–113, New York, NY, USA, 2010. ACM.

[31] H.-W. Yoo, D.-S. Jang, S.-H. Jung, J.-H. Park, and K.-S. Song. Visual information retrieval system via content-based approach. *Pattern Recognition*, 35(3):749–769, 2002.

[32] M. Zaffalon and M. Hutter. Robust feature selection by mutual information distributions. In *18th International Conference on Uncertainty in Artificial Intelligence*, pages 577–584, 2002.

[33] X. Zhang, V. Ramani, Z. Long, Y. Zeng, A. Ganapathiraju, , and J. Picone. Scenic beauty estimation using independent component analysis and support vector machines. In *Proceedings of IEEE Southeastcon*, pages 97–101, 1999.