Predicting Image Memorability by Multi-view Adaptive Regression

Houwen Peng[1,2]*, Kai Li[1]*, Bing Li[1], Haibin Ling[2], Weihua Xiong[1], Weiming Hu[1]
[1]Institute of Automation, Chinese Academy of Sciences
[2]Department of Computer & Information Sciences, Temple University
{houwen.peng, kai.li, bli, wmhu}@nlpr.ia.ac.cn hbling@temple.edu

ABSTRACT

The images we encounter throughout our lives make different impressions on us: Some are remembered at first glance, while others are forgotten. This phenomenon is caused by the intrinsic memorability of images revealed by recent studies [5, 6]. In this paper, we address the issue of automatically estimating the memorability of images by proposing a novel *multi-view adaptive regression* (MAR) model. The MAR model provides an effective mapping of visual features to memorability scores by taking advantage of robust feature selection and multiple feature integration. It consists of three major components: an adaptive loss function, an adaptive regularization and a multi-view modeling strategy. Moreover, we design an alternating direction method (ADM) optimization algorithm to solve the proposed objective function. Experimental results on the MIT benchmark dataset show the superiority of the proposed model compared with existing image memorability prediction methods.

Categories and Subject Descriptors

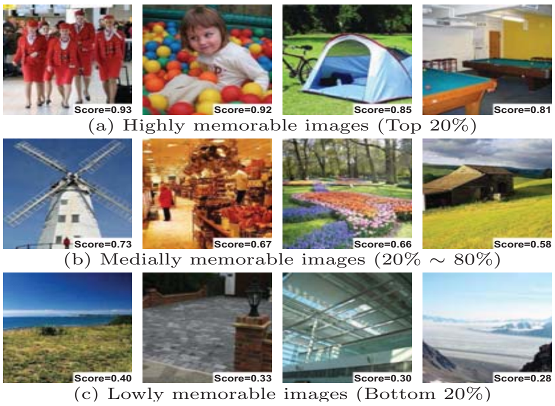
I.5.1 [**Pattern Recognition**]: Models—*statistical*; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Modeling and recovery of physical attributes*

General Terms

Algorithms, Experimentation

Keywords

Image Memorability, Adaptive Regression, Multi-view Learning, Prediction



(a) Highly memorable images (Top 20%)

(b) Medially memorable images (20% ∼ 80%)

(c) Lowly memorable images (Bottom 20%)

Figure 1: Sample images whose memorability scores are predicted by the proposed MAR method.

1. INTRODUCTION

Every day, we continuously encounter new photographs and images on social networks and in the media. While we may glance at them only once, some pictures stick in our minds whereas others fade away. Images are differentially memorable — not all are equal in our memory. This phenomenon demonstrates that memorability is an inherent property of individual images. It characterizes the probability that an observer will correctly recall a photograph after a period of time [6]. Predicting image memorability recently attracts lots of researchers' attention due to its promising applications in selecting magazine covers, designing logos, decorating websites and much more.

The prediction of an image's memorability is essentially a regression problem which maps an image (or its features) to a memorability score, where higher score indicates high memorability. Most existing prediction models are mainly based on support vector machine (SVM) technique. For example, Isola et al. [6, 5] propose to train support vector regression (SVR) on visual features and attributes to score the memorability of images. Khosla et al. [7] design a novel probabilistic process of memory forgetting, and exploit ranking SVM (RSVM) to sort the memorability of all images. To make use of attention mechanism, Celikkale et al. [2] introduce an attention-driven spatial pooling strategy for feature encoding and adopt SVR for prediction.

Although these existing SVM-based methods produce encouraging performance, two issues may prevent them from further improvement: (1) They combine multiple features through simple stacking or multiplication [6, 2], and feed them into regressors. Such combination schemes may not capture well the physical meaning of each feature and may cause information redundancy in learning. (2) They have

*H. Peng and K. Li contributed equally to this work.

MM'15, October 26–30, 2015, Brisbane, Australia.
© 2015 ACM. ISBN 978-1-4503-3459-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/XXX.XXXXXXX.

limitations in automatic feature selection, which is crucial for image memorability prediction since it remains an open problem that what makes an image memorable .

To address these problems, we propose a novel multi-view adaptive regression (MAR) model, which consists of three components: (1) an adaptive loss function which smoothly interpolates between the traditional ℓ_1 and ℓ_2-norm, and is robust to noises and outliers, (2) an adaptive penalty term which automatically selects desirable features according to training data, and (3) a multi-view framework which regards each visual feature as an individual view for memorability prediction. The proposed model not only pursues a meaningful combination of multiple features, but also takes advantage of complementary characteristics of these features from multiple views, e.g., low-level gradient information and high-level object semantic. Compared with existing memorability prediction models, the proposed method is more effective and robust thanks to its adaptive feature selection and multi-view modeling.

2. MULTI-VIEW ADAPTIVE REGRESSION

For clarity, before presenting the proposed multi-view adaptive regression, we first give out the single-view case, namely the adaptive regression model.

2.1 Adaptive Regression (AR)

Given a feature vector $\mathbf{x} \in \mathbb{R}^p$ representing an image, the prediction of its memorability score $y \in \mathbb{R}$ is a standard regression problem formulated as $f(\mathbf{x}) \rightarrow y$, where f is the fitting model learned on the training samples. In this paper, we consider a linear regression model f formulated as follows:

$$y = \mathbf{w}^T\mathbf{x} + \varepsilon, \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^p$ is the learned model's parameter, and ε is the residual error between the linear prediction and the true response. Given n training images with their feature representation $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n\times p}$ as well as corresponding memorability scores $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$, a common way to estimate the parameter vector $\mathbf{w}$ is to penalize the empirical risk minimization, which is defined as

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} L(\mathbf{y}, \mathbf{X}\mathbf{w}) + \lambda R(\mathbf{w}), \tag{2}$$

where $L(\cdot)$ is a loss function used to measure the error between the model's prediction and ground truth, $R(\cdot)$ is a regularization term to avoid overfitting through constraining the complexity of the model, and $\lambda > 0$ is a tradeoff parameter between the two items.

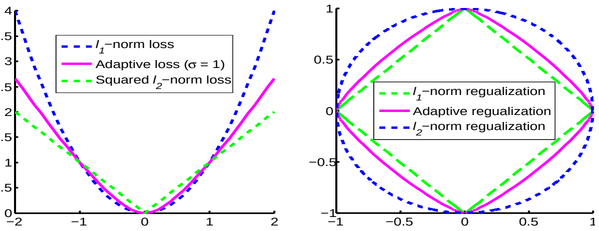
There exists a large body of loss functions and regularization alternatives in the community of statistics and machine learning. For the sake of stable feature selection and robust prediction, we propose a novel *adaptive regression* model which is composed of an adaptive loss and an adaptive regularization and formulated as:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_\sigma + \lambda\|\mathbf{X}\mathrm{Diag}(\mathbf{w})\|_*. \tag{3}$$

Here, $\|\cdot\|_\sigma$ is the adaptive loss function with parameter σ computed as

$$\|\mathbf{a}\|_\sigma = \sum_{i=1}^{n} \frac{(1+\sigma)a_i^2}{|a_i| + \sigma}, \quad \text{where} \quad \mathbf{a} = [a_1, \ldots a_n]^T, \tag{4}$$

$\|\cdot\|_*$ is the nuclear norm (the sum of singular values of a matrix), and Diag($\mathbf{w}$) converts the vector $\mathbf{w}$ into a diagonal matrix (the i-th diagonal entry is w_i). The adaptive regularization involves the sample matrix $\mathbf{X}$ and adaptively selects desirable feature subset for regression analysis according to the correlation information among samples [11].



(a) Loss (b) Regularization

Figure 2: Illustration of the adaptive loss and regularization, both of which are between ℓ_1 and ℓ_2-norm.

The proposed regression model is adaptive because: (1) The loss function $\|\cdot\|_\sigma$ takes advantage of both ℓ_1-norm loss and squared ℓ_2-norm loss by smoothly interpolating between them as shown in Fig. 2(a). Thus it is robust to the data outliers (under Laplacian distribution) and efficient in learning the normal data (under Gaussian distribution) [12]. (2) The adaptive regularization, a.k.a. trace lasso [4], balances the ℓ_1 and ℓ_2-norm according to input samples (see Fig. 2(b)), and simultaneously groups correlated data together and performs automatic feature selection.

2.2 Multi-view Adaptive Regression (MAR)

Exploiting information from multiple sources can effectively improve the prediction performance as a result of their complementary characteristics. In this subsection, we extend the proposed adaptive regression model to multi-view setting that takes multiple visual features into account.

Contrary to the existing methods which simply concatenate multiple feature vectors into a single one [6, 2], our method jointly combines all residuals, which are derived from each individual features independently, to minimize the loss in learning via a weight β. Thus, the objective function Eq.(3) is extended to multi-view setting:

$$\begin{aligned} &\min_{\mathbf{w}^v, \beta^v} \sum_{v=1}^{M} \|\mathbf{y} - \beta^v\mathbf{X}^v\mathbf{w}^v\|_\sigma + \lambda\|\mathbf{X}^v\mathrm{Diag}(\mathbf{w}^v)\|_*, \\ &s.t. \sum_{v=1}^{M} \beta^v{=}1, \quad \beta^v \geq 0. \end{aligned} \tag{5}$$

Here, $v \in \{1, ..., M\}$ indexes M types of features including color, texture, gradient, shape and semantics to be illustrated in Sec. 3. $\mathbf{X}^v$, $\mathbf{w}^v$ and β^v represent the feature matrix, model parameter and view weight corresponding to the v-th visual feature respectively. This multi-view extension not only preserves physical meaning of each feature, and also leverages their complementary characteristics for prediction.

2.3 Optimization

Considering the balance between efficiency and accuracy in practical applications, we adopt the well established alternating direction method (ADM) [10] to optimize the convex problem Eq.(5). (Eq.(3) can be viewed as a special case of Eq.(5)). We first introduce an auxiliary variable $\mathbf{H}^v$ to make

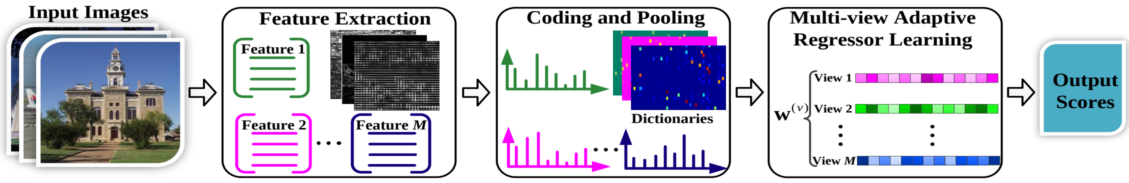


Figure 3: A flowchart to illustrate the proposed multi-view model for image memorability prediction.

the objective function Eq.(5) separable,

$$\begin{aligned} &\min_{\mathbf{w}^v,\beta^v} \sum_{v=1}^{M} \|\mathbf{y} - \beta^v \mathbf{X}^v \mathbf{w}^v\|_\sigma + \lambda \|\mathbf{H}^v\|_*, \\ &s.t.\ \ \mathbf{H}^v = \mathbf{X}^v \text{Diag}(\mathbf{w}^v),\ \sum_{v=1}^{M} \beta^v{=}1,\ \beta^v \geq 0. \end{aligned} \tag{6}$$

Now the problem Eq.(6) can be solved with the ADM, which minimizes the following augmented Lagrangian function:

$$\begin{aligned} &L(\mathbf{H}^v, \mathbf{w}^v, \beta^v) = \sum_{v=1}^{M} \|\mathbf{y} - \beta^v \mathbf{X}^v \mathbf{w}^v\|_\sigma + \lambda \|\mathbf{H}^v\|_* \\ &+ tr((\mathbf{\Phi}^v)^T(\mathbf{H}^v - \mathbf{X}^v \text{Diag}(\mathbf{w}^v))) + \phi^v(\sum_{v=1}^{M} \beta^v - 1) \\ &+ \tfrac{\mu}{2}\{\|\mathbf{H}^v - \mathbf{X}^v \text{Diag}(\mathbf{w}^v)\|_F^2 + (\sum_{v=1}^{M} \beta^v - 1)^2\}, \end{aligned} \tag{7}$$

where $\mathbf{\Phi}^v \in \mathbb{R}^{n\times p}$ and $\phi^v \in \mathbb{R}$ are the Lagrange multipliers, and $\mu > 0$ is the penalty parameter for violation of the linear constraints. To solve Eq.(7), we search for the optimal $\mathbf{H}^v$, $\mathbf{w}^v$ and β^v iteratively as summarized in Algorithm 1. Specifically, Step 1 in Alg.1 can be optimized by the singular value thresholding operator [1], while Step 2 is solved via the iteratively re-weighted algorithm proposed in [12].

3. MAR-BASED IMAGE MEMORABILITY PREDICTION

Algorithm 1 Solving MAR via ADM.

```
Input: $M$ types of feature matrices $\{\mathbf{X}^v\}_{v=1}^M \in \mathbb{R}^{n\times p}$, the memorability score $\mathbf{y} \in \mathbb{R}^n$, the parameters $\sigma$ and $\lambda$ .
Initialize: $\mathbf{H}^v, \mathbf{w}^v, \beta^v, \phi^v, \mathbf{\Phi}^v, \rho$, and $\mu_{\max}$.
    While not converged do
    1. Fix the others and update $\{\mathbf{H}^v\}_{v=1}^M$ by $\mathbf{H}^v = \arg\min_{\mathbf{H}^v} \frac{\lambda}{\mu}\|\mathbf{H}^v\|_* + \frac{1}{2}\left\|\mathbf{H}^v - (\mathbf{X}^v \text{Diag}(\mathbf{w}^v) - \frac{1}{\mu}\mathbf{\Phi}^v)\right\|_F^2$.
    2. Fix the others and update $\{\mathbf{w}^v\}_{v=1}^M$ by $\mathbf{w}^v = \arg\min_{\mathbf{w}^v} \|\mathbf{y} - \beta^v \mathbf{X}^v \mathbf{w}^v\|_\sigma + \frac{\mu}{2}\left\|\mathbf{X}^v \text{Diag}(\mathbf{w}^v) - (\frac{1}{2}\mathbf{H}^v + \frac{1}{\mu}\mathbf{\Phi}^v)\right\|_F^2$.
    3. Fix the others and update $\{\beta^v\}_{v=1}^M$ by $\beta^v = ((\mathbf{X}^v\mathbf{w}^v)^T\mathbf{D}^v\mathbf{X}^v\mathbf{w}^v + u/2)^{-1}[(\mathbf{X}^v\mathbf{w}^v)^T\mathbf{D}^v\mathbf{y} + (\mu - \phi^v)/2]$
       where $\mathbf{D}^v = diag(d_1^v, ..., d_n^v)$ and
       $d_i^v = (1+\sigma)\frac{\left\|y_i - \beta^v(\mathbf{w}^{(v)})^T\mathbf{x}_i^{(v)}\right\|_2 + 2\sigma}{2(\left\|y_i - \beta^v(\mathbf{w}^{(v)})^T\mathbf{x}_i^{(v)}\right\|_2 + \sigma)^2}$.
    4. Update the multipliers $\phi^v = \phi^v + \mu^v(\sum_{v=1}^{M} \beta^v - 1)$,
       $\mathbf{\Phi}^v = \mathbf{\Phi}^v + \mu^v(\mathbf{H}^v - \mathbf{X}^v \text{Diag}(\mathbf{w}^v))$.
    5. Update the parameter $\mu^v = \min(\rho\mu^v, \mu_{\max})$.
    6. Check the convergence conditions $\mathbf{H}^v - \mathbf{X}^v \text{Diag}(\mathbf{w}^v) \rightarrow 0$ and $\sum_{v=1}^{M} \beta^v - 1 \rightarrow 0$ for $v = 1, ..., M$.
    End While
Output: The coefficients $\{\mathbf{w}^v\}_{v=1}^M$ and $\{\beta^v\}_{v=1}^M$.
```

This section elaborates on the prediction of image memorability using the proposed multi-view adaptive regression model. Our prediction method consists of three major stages as illustrated in Fig. 3. (1) We first extract features from each input image considering color, texture, gradient, etc. (2) For each type of feature, we build a dictionary using k-means and apply local-constraint linear coding (LLC) [16] to soft-encode each feature into some dictionary entries. Similar to [5], we perform max pooling with a spatial pyramid matching (SPM) [16] to obtain the final feature vector in each view of the input image. (3) Finally, we exploit the proposed MAR model to learn a regressor on the training data, and then conduct prediction on the testing data.

In more details, we extract five common features in terms of low, middle and high-level visual information to represent images. Considering the power of low-level features in human vision system, we extract the color, texture and gradient for our task. For color feature, we convert the image to color names [15], then learn a dictionary of size 128 and apply LLC at 2-level SPM to obtain the color descriptor. To encode visual texture perception information, we use the popular local binary pattern (LBP) [13] and perform a 2-level SPM of non-uniform LBP descriptors. For gradient information, we densely sample HOG [3] with a cell size of 2×2 and build a dictionary of size 256. The descriptors are max-pooled at 2-level SPM using LLC. We further exploit the mid-level shape feature to represent images. The shape is denoted as a histogram of local self-similarity geometric patterns (SSIM [14]) with the size of 256 pooled at 2-level SPM. Moreover, high-level semantic meaning has been verified to be strongly correlated to image memorability [6]. Similar to [7], we use the automatic object bank [9] feature to model the presence of various objects in the images.

4. EXPERIMENTS AND RESULTS

4.1 Experimental Setup

We use the MIT image memorability dataset [6] to evaluate the proposed prediction model. This dataset contains 2222 natural images associated with human-annotated memorability scores. For the quantitative analysis, we use Spearman's rank correlation ρ and the precision-recall curve introduced in [6] to measure the performance of models. Same as [6], we evaluate the performance over 25 random splits of the dataset with an equal number of images for training and testing (1111). These train and test splits have been scored by different halves of the participants, showing a human consistency of $\rho = 0.75$, which can be viewed as an upper bound in the performance of prediction methods.

The tradeoff parameter λ of the MAR model is tuned on a validation set using 5-fold cross validation and set to be 1 to balance the adaptive regularization and loss function.

	Color	LBP	HOG	SSIM	Semantic	All (**MAR**)	SVR-MGF[6]	RSVM[7]	SVR-SO[2]	SVR-WOA[8]	Human
Top 20	76%	79%	83%	82%	82%	85%	83%	85%	84%	85%	86%
Top 100	73%	76%	80%	79%	80%	83%	80%	81%	81%	81%	84%
Bottom 100	60%	57%	56%	59%	58%	56%	56%	55%	56%	55%	47%
Bottom 20	57%	55%	54%	53%	53%	51%	54%	52%	55%	52%	40%
ρ	0.26	0.37	0.45	0.44	0.46	**0.52**	0.46	0.50	0.47	0.49	0.75

Table 1: Comparison of predictions. Left: results produced by our method considering each individual features and their combination. Right: baselines and the human prediction.

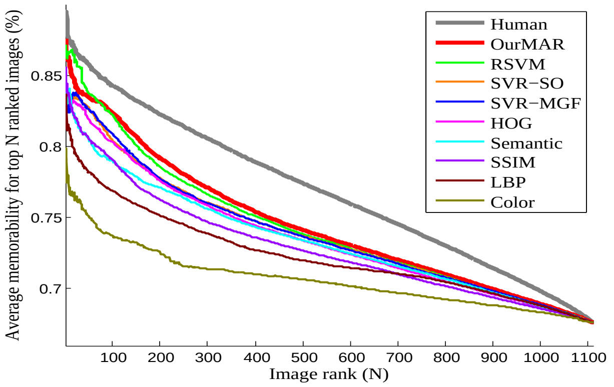


Figure 4: Comparison of precision-recall curves averaged across 25 random splits.

Upon our preliminary experiments, we empirically set the adaptive parameter as σ=0.1 in all experiments.

4.2 Results

We first compare the prediction performance of our MAR model with 4 other existing image memorability prediction methods on the MIT dataset. The baseline methods are SVR-MGF [6] which stacks multiple global features and uses SVR as the prediction model, RSVM [7] which fuses the results of ranking SVM learned on six common features independently, SVR-SO [2] which trains SVR over the combination of three groups of image features considering saliency and objectness, and SVR-WOA [8] which trains SVR on weighted object area features. Fig. 4 and Tab. 1 summarize the experimental results. It is observed that our method (MAR) outperforms SVR-MGF, SVR-SO and SVR-WOA by 10.8%, 8.7% and 4.3% respectively (see Tab. 1). This indicates the proposed multi-view learning model is more robust than the stereotyped regression methods at leveraging multiple features for image memorability prediction. Furthermore, our method achieves comparable performance compared to RSVM, and the average measured memorability of the 100 highest predicted images ("Top-100" shown in Tab. 1) of our method is superior. This may be caused by the fact that RSVM performs prediction by exploiting the ranking scheme, rather than the regression strategy used in our model. As shown in Fig. 4, the precision-recall curve of our method is also superior to others. Fig. 1 shows some sample images with their memorability scores predicted by our MAR method.

5. CONCLUSIONS

Predicting image memorability is a recent research topic which is crucial for the task of creating an image that a viewer will remember. In this paper, we propose a novel adaptive regression model to estimate image memorability. It is made up of an adaptive loss and an adaptive regularizer, and capable of selecting desirable features and robust to outliers. Moreover, we extend the regression model to multi-view case which effectively leverage multiple features for prediction. Extensive experiments show the superiority of our method compared with the state of the arts.

6. ACKNOWLEDGMENTS

This work is partly supported by the 973 basic research program of China (Grant No. 2014CB349303), NSFC (Grant No. 61472421, 61370038, 61202327, 61472227 and 61303086), the National 863 High-Tech R&D Program of China (Grant No. 2012AA012504), and the US NSF Grants US NSF Grants 1449860, IIS-1407156 and 1350521.

7. REFERENCES

[1] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):11, 2011.

[2] B. Celikkale, A. Erdem, and E. Erdem. Visual attention driven spatial pooling for image memorability. In *CVPRW*, pages 976–983, 2013.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[4] E. Grave, G. Obozinski, and F. Bach. Trace lasso: a trace norm regularization for correlated designs. In *NIPS*, 2011.

[5] P. Isola, D. Parikh, A. Torralba, and A. Oliva. Understanding the intrinsic memorability of images. In *NIPS*, pages 2429–2437, 2011.

[6] P. Isola, J. Xiao, A. Khosla, A. Torralba, and A. Oliva. What makes a photograph memorable? *IEEE TPAMI*, 2014.

[7] A. Khosla, J. Xiao, A. Torralba, and A. Oliva. Memorability of image regions. In *NIPS*, 2012.

[8] J. Kim, S. Yoon, and V. Pavlovic. Relative spatial features for image memorability. In *ACM Multimedia*, pages 761–764, 2013.

[9] L.-J. Li, H. Su, Y. Lim, and F.-F. Li. Object bank: An object-level image representation for high-level visual recognition. *IJCV*, 107(1), 2014.

[10] Z. Lin, R. Liu, and Z. Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *NIPS*, pages 612–620, 2011.

[11] C.-Y. Lu, J. Feng, Z. Lin, and S. Yan. Correlation adaptive subspace segmentation by trace lasso. In *ICCV*, 2013.

[12] F. Nie, H. Wang, H. Huang, and C. H. Q. Ding. Adaptive loss minimization for semi-supervised elastic embedding. In *IJCAI*, pages 1565–1571, 2013.

[13] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE TPAMI*, 24(7):971–987, 2002.

[14] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *CVPR*, 2007.

[15] J. van de Weijer, C. Schmid, and J. J. Verbeek. Learning color names from real-world images. In *CVPR*, 2007.

[16] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, pages 3360–3367, 2010.