# Pedestrian Attribute Recognition At Far Distance

Yubin Deng, Ping Luo, Chen Change Loy, Xiaoou Tang
Dept. of Information Engineering
The Chinese University of Hong Kong
danny.s.deng.ds@gmail.com, lp011@ie.cuhk.edu.hk, ccloy@ie.cuhk.edu.hk,
xtang@ie.cuhk.edu.hk

## ABSTRACT

The capability of recognizing pedestrian attributes, such as gender and clothing style, at far distance, is of practical interest in far-view surveillance scenarios where face and body close-shots are hardly available. We make two contributions in this paper. First, we release a new pedestrian attribute dataset, which is by far the largest and most diverse of its kind. We show that the large-scale dataset facilitates the learning of robust attribute detectors with good generalization performance. Second, we present the benchmark performance by SVM-based method and propose an alternative approach that exploits context of neighboring pedestrian images for improved attribute inference.

## Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Image databases

## Keywords

Large-scale database; attribute classification.

## 1. INTRODUCTION

Visual recognition of pedestrian attributes, such as gender, age, clothing style, is an emerging research topic in computer vision research, due to its high application potential in areas such as video-based business intelligence [16] and visual surveillance [5]. In many real-world surveillance scenarios, clear close-shots of face and body regions are not available. Thus, attribute recognition has to be performed at far distance using full body appearance (which can be partially occluded) in the absence of critical face/close-shot body visual information.

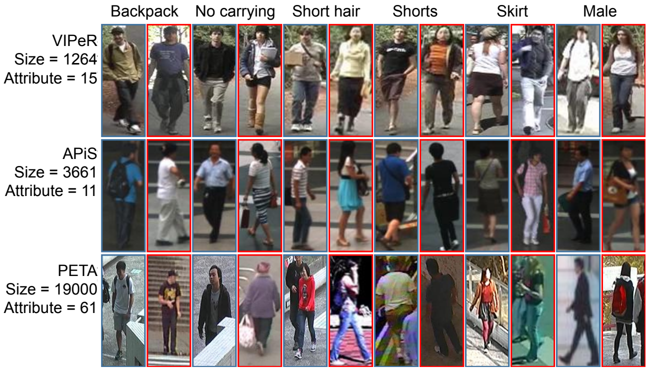


**Figure 1: The new PETA dataset contains far more images and attribute annotations than existing datasets. Positive and negative sample images are indicated by blue and red boxes, respectively.**

There are two fundamental challenges in attribute inference at far distance: 1) *Appearance diversity* - owing to diverse appearances of pedestrian clothing and uncontrollable multi-factor variations such as illumination and camera viewing angle, there exist large intra-class variations among different images for the same attribute. Learning to detect such attributes requires a rich set of training samples. Relying on a single source and small-scale training data would easily lead to an unrealistic model that generalizes poorly to unknown domains due to the inherent data bias. 2) *Appearance ambiguity* - far-view attribute recognition is an exceptionally difficult task due to inherent visual ambiguity and poor quality of visual features obtained from far view field (Fig. 1). In particular, an individual image may only occupy a few tens of imagery pixels whilst only a tiny fraction of them are truly distinctive for attribute classification. Often, parts of the body are occluded, either by obstacles or other pedestrians, which further increases the difficulty of extracting relevant features for inference. For instance, images with the 'carrying backpack' attribute may not necessarily have the full bag visible due to pedestrian posture (Fig. 1).

Existing datasets do not reflect the diversity nature in real-world environment. In view of this shortcoming, we introduce and release a new large-scale PEdesTrian Attribute (PETA) dataset[1]. The dataset is by far the largest of its kind, covering more than 60 attributes on 19000 images. As can be seen from Fig. 1, in comparison with existing datasets, PETA is more diverse and challenging in terms of imagery variations and complexity. More details are presented in Sec. 2. Apart from releasing the new dataset, we also propose an alternative approach for attribute recogni-

*MM'14,* November 3–7, 2014, Orlando, Florida, USA.
Copyright 2014 ACM 978-1-4503-3063-3/14/11 ...$15.00.
http://dx.doi.org/10.1145/2647868.2654966.

[1] http://mmlab.ie.cuhk.edu.hk/projects/PETA.html

tion with emphasis to mitigating the visual ambiguity of appearance features. Specifically, instead of treating an image independently, we consider inference with the help from neighboring pedestrian images whose appearances look alike. We hypothesize that neighboring samples share natural invariance in their feature space, which could be treated as a form of regularization or context. As such, attribute inference of an image can be locally constrained by its neighbors to obtain a more reliable prediction. To this end, we view multiple pedestrian images as forming a Markov Random Field (MRF) graph. The underlying graph topology is automatically inferred, with node associations weighted by pairwise image similarity. The similarity can be estimated as the conventional Euclidean distance or more elaborated decision forest-based similarity with feature selection [20, 21]. By carrying out inference on the graph, we jointly reason and estimate the attribute probability of all images in the graph.

It is worth noting that MRF inference for smoothing [9] is commonly applied in image segmentation [15], but this paper is the first work that explores this approach for pedestrian attribute inference. We summarize our contributions as follows: 1) we introduce the largest pedestrian attribute dataset to date to facilitate future research on attribute classification at far distance; 2) we present the benchmark performance by SVM-based method [8] and propose an alternative approach that exploits context of neighboring images for improved attribute inference. Thanks to the neighboring context, our model is capable of accurately detecting subtle attributes, which may otherwise be mis-detected from single image.

## 2. PEDESTRIAN ATTRIBUTE DATASET

**Statistics**: We carefully chose and organized 10 publicly available small-scale datasets to construct the new PEdesTrian Attribute (PETA) dataset[2,3]. The name, size, and main characteristics of constituent datasets, are summarized in the table in Fig. 2. The PETA dataset thus consists of 19000 images, with resolution ranging from 17 × 39 to 169 × 365 pixels. Organizing these datasets are not straightforward. First, we carefully removed erroneous images or duplicated copies from each datasets. In addition, each image is newly labeled with 61 binary and 4 multi-class attributes. The binary attributes cover an exhaustive set of characteristics of interest, including demographics (e.g. gender and age range), appearance (e.g. hair style), upper and lower body clothing style (e.g. casual or formal), and accessories. The four multi-class attributes encompass 11 basic color namings [17], respectively, for footwear, hair, upper-body clothing, and lower-body clothing. The distribution of a binary attribute is considered balanced if the ratio of larger to smaller class is no more than 20:1. As such, out of the 61 binary attributes, 31 are balanced. Fig. 2(b) depicts the distribution of a few attributes with sample images[4].

**Uniqueness**: Compared to existing pedestrian attribute dataset, this new attribute dataset has three notable uniqueness: (1) *Larger size*: the size of the PETA dataset is over 5× and 15× larger than the APiS and VIPeR datasets, respectively. (2) *High diversity*: we deliberately selected smaller-scale datasets collected under different conditions from diverse scenes to enrich the composition of the new attribute dataset. As can be seen from Fig. 1 and summarized in the table in Fig. 2(a), despite that the constituents of PETA are all captured from far view field, they exhibit large differences in terms of lighting condition, camera viewing angles, image resolutions, background complexity, and indoor/outdoor environments. (3) *Rich annotations*: The PETA dataset contains far richer annotations in comparison with existing datasets, such as VIPeR [6], with only 15 binary attributes, and APiS [19], with 11 binary and 2 multi-class attributes, It is worth pointing out that the 61 annotated attributes in PETA dataset include the 15 attributes that are suggested by the UK Home Office and UK police to be the most valuable in tracking and criminal identification [14].

**Usage**: Visual understanding through semantic attributes is an active research topic in the multimedia research community, e.g. [12], for the applications such as image retrieval and recommendation systems [3, 7, 11]. This dataset can serve as an alternative benchmark. More specifically, the new dataset can be used in visual surveillance research on pedestrian tracking, detection, re-identification, and activity analysis. Furthermore, the visual attributes can potentially be integrated with multi-sources of information, e.g. audio, textual annotation, sensor signals, for multimedia surveillance [20, 4]. In the next section, we present two benchmarking methods for attribute classification, a fundamental task in visual understanding.

## 3. BASELINE METHODS

**Baseline 1.** SVM with intersection kernel (ikSVM) [13][5] reduces both the time and space complexities from $\mathcal{O}(mn)$ of the traditional linear kernel SVM to $\mathcal{O}(n)$. Here, $m$ and $n$ are the number of the support vectors and the dimension of feature vectors. Previous study [8] has applied this method successfully for pedestrian attribute classification. Cross validation for slack parameter $C$ is performed as in [8].

**Baseline 2.** To improve attribute inference, we exploit the context of neighboring images by Markov Random Field (MRF), which is an undirect graph, where each node represents a random variable and each edge represents the relation between two connected nodes. The energy function of MRF over a graph $G$ can be defined as follows

$$E_{MRF}(G) = \sum_{u \in G} C_u(l_u) + \sum_{u \in G} \sum_{v \in N(u)} S_{uv}(l_u, l_v), \qquad (1)$$

where $u, v \in G$ are two random variables in the graph and $l_u$ denotes the state of $u$. $C_u$ and $S_{uv}$ signify the unary cost and pairwise cost functions, respectively. More precisely, they indicate the cost of assigning state $l_u$ to variable $u$ as well as the cost of assigning states to neighboring nodes $u, v$, which is determined based on the graph structure (e.g., assigning different states to nodes that are similar is penalized). $N(u)$ is a set of variables that are the neighbors of $u$.

In this work, each random variable corresponds to an image and the relation between two variables corresponds

[2] Images in PETA dataset are all exclusive from those in APiS [19].

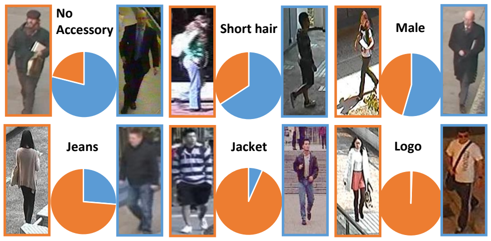
[3] All images in PETA are freely available for academic use except i-LIDS, which requires application to United Kingdom Home Office, https://www.gov.uk/imagery-library-for-intelligent-detection-systems.

[4] Sample images of the datasets and the full distributions of all attributes can found in the supplementary material.

[5] http://www.cs.berkeley.edu/~smaji/projects/fiksvm/

| Datasets | #Images | Camera angle | View point | Illumination | Resolution | Scene |
|---|---|---|---|---|---|---|
| 3DPeS | 1012 | high | varying | varying | from 31x100 to 236x178 | outdoor |
| CAVIAR4REID | 1220 | ground | varying | low | from 17x39 to 72x141 | outdoor |
| CUHK | 4563 | high | varying | varying | 80x160 | outdoor |
| GRID | 1275 | varying | frontal&back | low | from 29x67 to 169x365 | indoor |
| i-LIDS | 477 | medium | back | high | from 32x76 to 115x294 | indoor |
| MIT | 888 | ground | back | high | 64x128 | outdoor |
| PRID | 1134 | high | profile | low | 64x128 | outdoor |
| SARC3D | 200 | medium | varying | varying | from 54x187 to 150x307 | outdoor |
| TownCentre | 6967 | medium | varying | medium | from 44x109 to 148x332 | outdoor |
| VIPeR | 1264 | ground | varying | varying | 48x128 | outdoor |
| **Total = PETA** | **19000** | **varying** | **varying** | **varying** | **varying** | **varying** |

(a)



(b)

**Figure 2: (a) The composition of PETA dataset. (b) Examples of attributes with their distribution (blue: positive, orange: negative).**

to the similarity between images. The states of variable are the values of the image attribute, which is $l_u \in \{0, 1\}$. The unary function is modeled by

$$C_u(l_u) = -\log P(l_u|u), \tag{2}$$

where $P(l_u|u)$ is the probability of predicting the attribute value of image $u$ as $l_u$. This probability is learned by ikSVM.

Now we consider the definition of the pairwise function. To define affinity between nodes, a simple way widely adopted by existing methods, such as [18], is the Gaussian kernel, $\exp\{-\frac{\|u-v\|^2}{\sigma^2}\}$, in which $u, v$ indicate the feature vectors of two images and $\sigma$ is a coefficient that needs to be tuned. The graph built on this kernel function can model the global smoothness among images. However, when large variations are presented, one may consider modeling the local smoothness and discovering the intrinsic manifold of the data. Thus, an alternative is to employ the random forest (RF) [2] to learn the pairwise function [20, 21]. The RF we adopted is unsupervised, i.e. it takes unlabeled test samples as input. The output is pairwise sample similarity derived from the data partitioning discovered at the leaf nodes of RF. The unsupervised RF can be learned using the pseudo two-class method as in [20, 21, 10]. Specifically, we treat the original unlabeled test samples as first class. The pseudo second class is created by sampling at random from the univariate distributions of the unlabeled test samples. With this strategy, the unsupervised RF learning problem becomes a canonical classification problem that can be solved by conventional classification forest training method. Specifically, the information gain of unsupervised RF is identical to that of conventional supervised RF, defined as

$$\Delta\mathcal{I} = \mathcal{I}_p - \frac{n_l}{n_p}\mathcal{I}_l - \frac{n_r}{n_p}\mathcal{I}_r, \tag{3}$$

where $p, l$, and $r$ refer to a splitting node and its left and right child. The variable $n$ denotes the number of samples at a node, $n_p = n_r + n_l$. $\mathcal{I}$ is the Gini impurity measure at each node [2].

The pairwise function is expressed as

$$S_{uv}(l_u, l_v)) = \begin{cases} \frac{1}{T}\sum_{t=1}^{T}\exp\{-dist^t(u,v)\} & \text{if } l_u \neq l_v, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Here, $dist^t(u, v) = 0$ if $u, v$ fall into the same leaf node and $dist^t(u, v) = +\infty$ otherwise, where $t$ is the index of tree. Since the graph is dense, the inference of MRF is difficult. Thus, we build a $k$-NN sparse graph by limiting the number of neighbors for each node. We set $k = 5$ in our experiment. Eq.(1) can be efficiently solved by the min-cut/max-flow algorithm introduced in [1].

# 4. EXPERIMENT

We present benchmark results on PETA by evaluating the performance of intersection kernel SVM (ikSVM) [13], MRF with Gaussian kernel (MRFg), and MRF with random forest (MRFr), as discussed in Sec.3.

We randomly partitioned the dataset images into 9,500 for training, 1,900 for verification and 7,600 for testing. We selected 35 attributes for our study, consisting of the 15 most important attributes in video surveillance proposed by human experts [8], [14] and 20 difficult yet interesting attributes chosen by us, covering all body parts of the pedestrian and different prevalence of the attributes. For example, 'sunglasses' and 'v-neck' have a limited number of positive examples. For the attributes with unbalanced positives and negatives samples, we trained ikSVM for each attribute by augmenting the positive training examples to the same size as negative examples with small variations in scale and orientation. This is to avoid bias due to imbalanced data. For MRFg and MRFr, we built the graphs using two different schemes. The first scheme, symbolized by MRFg$_1$ and MRFr$_1$, is to construct the graphs with only the testing images. The second one, symbolized by MRFg$_2$ and MRFr$_2$, is to include both training and testing samples in the graphs.

**Features.** Low-level color and texture features have been proven robust in describing pedestrian images [8], including 8 color channels such as RGB, HSV, and YCbCr, and 21 texture channels obtained by the Gabor and Schmid filters on the luminance channel. The setting of the parameters of the Gabor and Schmid filters are given in [8]. We horizontally partitioned the image into six strips and then extracted the above feature channels, each of which is described by a bin-size of 16. Finally, a 2784-dimensional feature vector is obtained for each image.

As shown in Table 1, we report attribute detection accuracy as [8] and have the following observations. First, the MRF-based methods outperform ikSVM on most of the attributes. For instance, MRFr$_2$ achieves an average of 3.4% improvement over ikSVM for the 'age' attributes shown on the top. This is significant in a dataset with large appearance diversity and ambiguity and it demonstrates that graph regularization can improve attribute inference. Second, the MRF graphs built with the second scheme is superior compared to the first scheme. This is reasonable because using both the training and testing data can better cover the image space. Third, for many important attributes, such as 'hair' and 'gender', random forest works much better than Gaussian kernel. It is worth pointing out that all methods perform poorly on attributes with imbalanced positive-negative distribution, such as 'logo', 'sandals', 'sunglasses', and 'v-neck'.

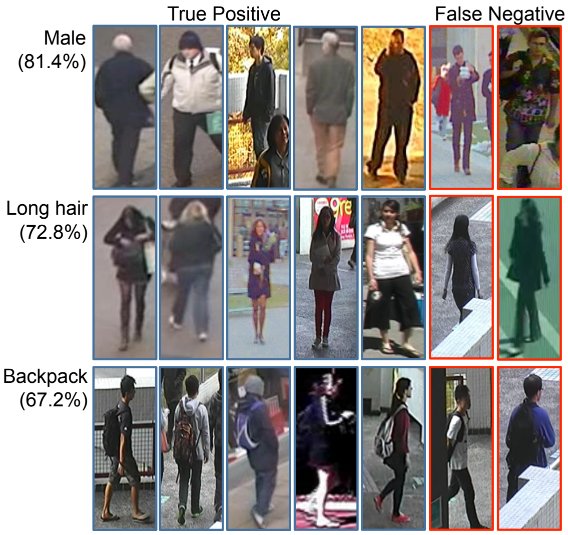


**Figure 3: Examples of attribute classification.**

Fig. 3 shows some attribute classification results using forest MRF. By exploiting the hidden context information, we obtain good accuracy on gender and hair attributes. False negative samples typically result from occlusion (e.g. backpack), color ambiguity (long hair) and background noise (male).

## 5. CONCLUSIONS

This paper presents a new large-scale dataset (PETA) sized at 19000 images with 61 annotated attributes. To cope with such large and diverse data in the context of attribute classification, we explored and proposed a novel approach by exploiting the neighborhood information among image samples. We showed that accurate attribute detection can be achieved with the automatically inferred neighborhood graph topology. Future work will involve the evaluation of other popular algorithms in attribute classification using our dataset. We hope that this dataset could serve as a new benchmark for the more realistic training and testing of algorithms.

**Table 1: Accuracy on the PETA dataset.**

| Attribute | ikSVM | MRFg1 | MRFg2 | MRFr1 | MRFr2 |
|---|---|---|---|---|---|
| Age16-30 | 80.4 | 80.9 | 81.7 | 80.9 | **83.8** |
| Age31-45 | 73.6 | 74.6 | 76.2 | 74.0 | **78.8** |
| Age46-60 | 73.1 | 74.1 | 75.2 | 73.2 | **76.4** |
| AgeAbove61 | 87.2 | 87.2 | 88.2 | 86.3 | **89.0** |
| Backpack | 66.7 | 67.1 | 67.1 | 67.0 | **67.2** |
| CarryingOther | 64.6 | 64.9 | 66.8 | 64.6 | **68.0** |
| Casual lower | 70.7 | 70.9 | **71.6** | 70.4 | 71.3 |
| Casual upper | 70.3 | 70.4 | 71.2 | 69.8 | **71.3** |
| Formal lower | 71.0 | 71.2 | 71.8 | 71.2 | **71.9** |
| Formal upper | 70.0 | 70.3 | **70.4** | 70.3 | 70.0 |
| Hat | 82.3 | 82.9 | 84.3 | 82.3 | **86.7** |
| Jacket | 67.7 | 68.3 | **68.4** | 68.1 | 67.9 |
| Jeans | 74.9 | 75.2 | **76.1** | 75.0 | 76.0 |
| Leather shoes | 78.9 | 80.1 | 80.9 | 79.1 | **81.7** |
| Logo | **51.1** | **51.1** | **51.1** | **51.1** | 50.7 |
| Long hair | 71.5 | 71.7 | 72.6 | 71.8 | **72.8** |
| Male | 79.7 | 80.3 | 80.9 | 80.6 | **81.4** |
| MessengerBag | 71.8 | 72.9 | 74.3 | 72.7 | **75.5** |
| Muffler | 88.0 | 88.3 | 89.5 | 86.5 | **91.3** |
| No accessory | 76.8 | 77.2 | 78.6 | 77.1 | **80.0** |
| No carrying | 70.4 | 70.6 | **71.6** | 70.6 | 71.5 |
| Plaid | 64.0 | 64.5 | 64.5 | **65.0** | **65.0** |
| Plastic bag | 74.9 | 74.9 | 75.5 | 73.9 | **75.5** |
| Sandals | **50.3** | **50.3** | **50.3** | **50.3** | **50.3** |
| Shoes | 70.6 | 71.0 | 72.5 | 70.8 | **73.6** |
| Shorts | 56.0 | **56.5** | **56.5** | **56.5** | **56.5** |
| ShortSleeve | 71.3 | 71.7 | **71.8** | 71.7 | 71.6 |
| Skirt | 64.0 | 64.0 | 64.0 | 64.0 | **64.3** |
| Sneaker | 67.5 | 68.1 | 69.0 | 68.2 | **69.3** |
| Stripes | 51.5 | **52.3** | **52.3** | **52.3** | **52.3** |
| Sunglasses | **52.4** | **52.4** | **52.4** | 51.8 | 51.7 |
| Trousers | 74.0 | 74.5 | 75.7 | 75.7 | **76.5** |
| T-shirt | 64.3 | 64.5 | **64.6** | 63.6 | 64.2 |
| UpperOther | 80.7 | 80.7 | 81.8 | 81.1 | **83.9** |
| V-Neck | **51.1** | **51.1** | **51.1** | **51.1** | **51.1** |
| AVERAGE | 69.5 | 69.9 | 70.6 | 69.7 | **71.1** |

## 6. REFERENCES

[1] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *TPAMI*, 2004.

[2] L. Breiman. Random forests. *Machine learning*, 2001.

[3] J. Cai, Z.-J. Zha, W. Zhou, and Q. Tian. Attribute-assisted reranking for web image retrieval. In *MM*, pages 873–876. ACM, 2012.

[4] R. Cucchiara. Multimedia surveillance systems. In *Int. workshop on video surveillance & sensor networks*. ACM, 2005.

[5] S. Gong, M. Cristani, C. C. Loy, and T. Hospedales. The re-identification challenge. In S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors, *Person Re-Identification*. Springer, 2013.

[6] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE PETS*, 2007.

[7] Y. Han, F. Wu, X. Lu, Q. Tian, Y. Zhuang, and J. Luo. Correlated attribute transfer with multi-task graph-guided fusion. In *MM*. ACM, 2012.

[8] R. Layne, T. M. Hospedales, S. Gong, et al. Person re-identification by attributes. BMVC, 2012.

[9] S. Z. Li. *Markov random field modeling in computer vision*. Springer-Verlag New York, Inc., 1995.

[10] B. Liu, Y. Xia, and P. S. Yu. Clustering through decision tree construction. In *CIKM*. ACM, 2000.

[11] L. Liu, H. Xu, J. Xing, S. Liu, X. Zhou, and S. Yan. Wow! you are so beautiful today! ACM-MM, 2013.

[12] P. Luo, X. Wang, and X. Tang. A deep sum-product architecture for robust facial attributes analysis. ICCV, 2013.

[13] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. CVPR, 2008.

[14] T. Nortcliffe. People analysis cctv investigator handbook. *Home Office Centre of Applied Science and Technology*, 2011.

[15] C. Rother, T. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of image pairs by histogram matching-incorporating a global constraint into MRFs. CVPR, 2006.

[16] C. Shan, F. Porikli, T. Xiang, and S. Gong. *Video Analytics for Business Intelligence*, volume 409. Springer, 2012.

[17] J. Van De Weijer, C. Schmid, J. Verbeek, and D. Larlus. Learning color names for real-world applications. *TIP*, 18(7):1512–1523, 2009.

[18] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang. Joint multi-label multi-instance learning for image classification. CVPR, 2008.

[19] J. Zhu, S. Liao, Z. Lei, D. Yi, and S. Z. Li. Pedestrian attribute classification in surveillance: Database and evaluation. In *ICCV Workshop*, 2013.

[20] X. Zhu, C. C. Loy, and S. Gong. Video synopsis by heterogeneous multi-source correlation. In *ICCV*, 2013.

[21] X. Zhu, C. C. Loy, and S. Gong. Constructing robust affinity graph for spectral clustering. In *CVPR*, 2014.