RECENT PROGRESS IN ATTRIBUTES BASED LEARNING: A SURVEY

Hafeez Anwar, Martin Kampel
Computer Vision Lab
Institute of Computer Aided Automation
Vienna University of Technology, 1040
Vienna, Austria
email: hafeezcse@gmail.com, martin.kampel@tuwien.ac.at

Rehanullah Khan
Sarhad University of science and IT
Peshawar, KPK, Pakistan
email: rehanmarwat1@gmail.com

ABSTRACT
The recent use of attributes for various problems like object classification, object description and image retrieval produced promising results. The data sets are developed in which images are annotated with the attributes of objects instead of the category names. Attributes performed better than low-level features for cross-category generalization on several data sets. These data sets opened the avenues for potential research in computer vision and image understanding based on attributes. Therefore it is quite necessary to get an overview of the attributes, the way they work and data sets annotated with attributes. In this work we provide an overview of various solutions based on attributes, the available data sets that are annotated with the attributes and future directions for some potential problems where attributes may perform better than low-level features.

KEY WORDS
Attributes, object recognition, image retrieval

1. Introduction
Attributes are the visual properties of objects. They are mainly related to the appearance and geometrical structures of the objects[1]. Visual attributes used by humans to describe an object are called semantic attributes. They are used to represent parts, shape and materials of objects[2]. Humans have the ability to describe unfamiliar objects up to some extent by using the semantic attributes. An Object category recognition algorithm based on training images is likely to fail when presented with an image from a particular category for which no images are included in the training data set. It can be observed in the real world that the objects share visual and semantic attributes. Researchers have taken the advantage of the sharing nature of the attributes and have shown that attributes work well in situations where the images from a particular object category are absent in the training data set. Attributes have been used to identify an object or at least parts of an unknown object [2][3][4].

Apart from object category recognition, attributes can be used for image retrieval in a way that is natural and favorable for humans. Current methods of image retrieval use local or global features of images which are more related to the rigid structure of objects being queried. Semantic attributes can be used to describe an object due which they present a natural choice for image retrieval. Attributes have also been used for face verification[5], describing aesthetics and interestingness of images[6] and generation of sentences from images[7] as well.

The low level features are extracted from annotated training images and then quantized. The attributes classifiers are then learnt based on these quantized low-level features. As stated earlier, attributes are used to represent materials, parts and shapes. Color and texture are used for materials, visual words are used for parts, and edges are used for shapes[2]. Low-level feature used for these attributes are texture descriptors extracted with a textron filter bank, HOG[8] descriptor for the construction of visual words, and canny edge detectors[9] is used for edge detection. Apart from these, SIFT[10], rgSIFT[11], PHOG[12], SURF[13] and self-symmetry Histograms[14] are also used as low-level features. In most of the works, a classifier is trained for each attribute based on the quantized low-level features. The learned parameters from these classifiers are then used to predict the presence or absence of an attribute in a given image. Most of the algorithms learn these parameters by localizing the objects of interest in a bounding box that is pre-detected thus making the problem as “what is this” instead of “where is this”[2]. Some of the algorithms combine the localization as well as.

Attributes are the mid-level features which can be named and described making them more efficient than low level features for problems like object category recognition. They can efficiently model the visual characteristics of objects and their spatial relationships in an image. Attributes can also represent image properties and concepts. They have been shown to give better performance than low-level features in face verification [5] and image aesthetics prediction [6]. These results give the motivation to switch from solutions based on low-level features to attributes based solutions. Therefore an overview of the attributes and their applications becomes a pre-requisite before one can apply them to solution of a problem. This article will give the required overview.

Following are the contributions of this article.

• An overview of the attributes based solutions of object classification and description, image retrieval and image aesthetics prediction.

• Description of the attributes data sets.
• Potential future directions.

Rest of the paper is organized as follows. Section 2 summarizes the techniques that use attributes for object classification and description. An overview of the attributes used for image aesthetics prediction is given in section 3, section 4 gives an overview of attributes based image retrieval methods, the attributes data sets and future directions to apply attributes for some potential problems are given in section 6 and section 7 concludes the article.

2. Attributes based Classification

In attributes based classification a class or category is represented by its attributes. The attributes based class or category representation is more discriminative than example based representation due to the fact that the attributes are usually extracted using an expert knowledge. The example based approach is shown in Figure 1 while the attributes based approach is shown in Figure 2. In this section, methods based on attributes for object classification are reviewed. In these methods the objects are either pre-detected by bounding boxes or localized with the help of an object category detector. Attributes classifiers are then trained for the localized objects. The learned parameters of these classifiers are used to predict the attributes of objects in a given image. In most of the methods, the attributes are used for efficient transfer learning where the training and testing are done on images from different sets of classes.

2.1 Visual attributes extraction

Ferrari et al.[1] have used a probabilistic generative model to learn the visual attributes. These attributes are related to the appearance such as ‘color’ or ‘texture’ and shape such as ‘round’ or ‘cylindrical’. The basic units used for learning the visual attributes are segments of an image. A code book is constructed from the segments of the training images where each entry of the code book is a representative appearance descriptor of a subset of segments of the training images. A given image is divided into foreground and background segments, foreground segments being the ones containing a given visual attribute. Unary attributes, which are entirely characterized by the properties of a single segment, are either defined by appearance or by shape. The only observed variables for unary attributes are individual segments. The conditional probability for each segment is then calculated given the allowed appearance and shape. Binary attributes have basic elements composed of two segments like black and white stripes of a Zebra. For binary attributes, the image is represented as pairs of segments over all possible segments. The likelihood for appearance, shape and relative orientation and relative area is then calculated for each pair of segments. The model is learnt for both the unary as well as the binary attributes using a discriminative approach i.e. using a set of positive and negative images with respect to the visual attributes. The task of learning is to determine the model parameters that maximize a likelihood ratio. This learning gives the characteristic properties of an attribute and estimates its distribution. The area covered by an attribute is localized as well for the foreground segments.

2.2 Attributes based object description

The method proposed by Farhadi et al.[2] performs the object description based on semantic attributes. The base features are extracted from pre-localized objects in training data set. Base features consist of color and texture for materials, visual words for parts and edges for shapes. The undesirable attributes correlations occur during the learning stage of the attributes classifiers. To cope with this problem, a novel feature selection method is proposed. Their feature selection method focuses on within category prediction ability. Those features are selected which best distinguish object of the same category with and without a specific attribute. Then all the selected features over all the categories are pooled to learn a classifier for a specific attribute. These features are called selected features. Two data sets have been developed for this purpose called the a-Pascal and a-Yahoo data sets. These data sets are annotated with 64 semantic attributes using the Amazon's
Mechanical Turk [15]. Both the data sets have objects belonging to different categories yet the attributes assignment is done using across category prediction where training and test instances are drawn from different sets of classes. Apart from the attribute assignment, the proposed method also reports the absence of typical and presence of atypical attributes. The proposed method also works well at naming known objects, learning new categories with few visual examples and learning new categories based on pure textual descriptions.

Gang et al.[4] combine both the attributes and object category learning. In this method the objects are localized and their attributes are described. The training images are weakly labeled where every training image is annotated to contain an attribute-object pair but the location is unknown. Each image is a “bag” of windows at different scales and locations. If the image is labeled as positive then at least one window contains the object and labeled as negative if none of the windows contains the object. The attributes and category detectors are combined to simplify the problem of localization where both the detectors support each other at the location where an object (resp. attribute) is present. The main problem is the large number of candidate windows. The visual saliency [16] and homogeneity [17] are used to sub-sample the large set of windows in the bag to emphasize interesting image windows. The large bag of windows is thus reduced to small subset based on the scores obtained for saliency and homogeneity. The combined attributes and object category classifiers are then learnt using mi-SVM [18] which is an SVM for multiple instances learning with the constraint that both the classifiers give maximum responses in the windows where both the object and attribute are present.

The proposed method of Farhadi et al.[19] used the parts based model of [20] and category detectors to localize objects and then describe them by using the spatial relationship of the attributes. A new data set is constructed and annotated using the Amazon’s Mechanical Turk [15]. Objects are grouped into broader categories. These objects are localized and then attributes are assigned to them. Generalization is improved through efficient knowledge transfer giving better performance at localization and naming as well as inferring pose, composition and function of the objects. By learning one set of animals and vehicles, many others can be localized giving an improved generalization across broad domains. 28 types of objects (animals and vehicles) as well as several types of parts and ten types of materials are annotated. These annotations include object segmentation, object parts segmentation, category and parts labels, and masks for common materials, pose and viewpoint. The aim of the data set is to study the cross-category generalization with respect to localization and description. Detectors are trained on this data set for parts such as “wheels”, super ordinate categories such as “four-wheeled vehicle” and basic level categories such as “car” using the parts based and category based models. To localize objects in an image, these trained parts and category detectors are applied. Votes are accumulated from confident detectors to obtain object candidates. A graphical model is then used to describe the localized object.

2.3 Attributes based unseen object classification

Lampert et al.[3], used attributes for object classification where the training and the test sets are disjoint. In a usual object classification problem, the classifier is trained on hundreds of labeled images to detect future object instances. However if the future object instance belongs to a completely different category than those used in the training data set, the system fails to name it properly. In this work, a data set of 30,000 animal images from 50 categories is annotated with 85 attributes. The training and test sets are kept disjoint and two types of classifications based on semantic attributes have been proposed. Both the direct attributes prediction and indirect attributes prediction use a layer of attributes to decouple the samples and their corresponding labels. During training, the likelihoods of the attributes are learnt from the training samples where the attributes are represented in the form of a vector. The value of an attribute is ‘1’ if it is present in the sample and ‘0’ otherwise. The likelihoods of unknown classes are learnt using a minimal human effort. At test time, the posterior distribution of the training class labels induces a distribution over the labels of unseen classes by means of the class-attribute relationship.

2.4 Object recognition based on visual attributes extracted from text

Wang et al.[21] proposed to extract the visual attributes i.e. shape, color and pattern from pure textual descriptions given by an expert source. The proposed method builds ‘templates’ for the visual attributes of 10 species of butterflies. These templates contain slots for colors, patterns and their location on the butterfly. The techniques of Natural Language Processing are applied to the text extracted from an expert source i.e. eNature¹ to fill these templates. The properties of butterflies described by the expert source are both detailed and discriminative thus making the assignment of the attributes more efficient. In order to match the visual attributes of an image of a butterfly to the ones extracted from the text, the butterfly in the image is separated from the background using a semi-automatic segmentation proposed in [22]. Two visual attributes which are determined as salient by the textual description are a) dominant (Wing) color and b) colored spots. The spots are extracted using the Difference of Gaussian [10] and SIFT [11] descriptors are extracted around each candidate spot. A spot classifier is then trained to differentiate spots from non-spots using hand-marked butterfly images without incorporating the category information of the butterfly. The templates also give the color names of the wings and spots. For each color name in the template, a probability distribution is learnt from the training images using the L*a*b*

¹http://www.enature.com/fieldguides/
space.

A generative model is developed to predict the category of a given image of a butterfly. The priors over the dominant (wing) color and spot color are learnt from the templates. The likelihood of a given image is then evaluated for all the categories and the category that maximizes the likelihood is assigned to this image. Experiments are carried out on native and non-native English speakers as well. The description of a certain category is presented to them along with 10 images randomly selected from each category. They are told to select the image that best matches the category explained in the text. The same experiments are performed with the learnt model and the results obtained are comparable to the ones obtained from non-native English speakers.

### 3. Attributes based Image Aesthetics and Interestingness Prediction

The high level describable attributes are used by Dhar et al.[6] to select an image with high aesthetic quality from a large database. Describable attributes are those attributes of an image which can affect the human description of that particular image. The describable attributes used in this work to estimate the aesthetics and interestingness of an image are compositional attributes, content attributes and sky illumination attributes.

#### 3.1 Compositional attributes

Among the compositional attributes, the first one is the presence of salient objects. Three features related to saliency are used which are multi-scale contrast map, a center surround histogram map and center weighted color spatial distributions. These features are fed to a conditional random field (CRF) to predict the location of the salient object. The CRF is trained on positive and negative images with respect to a salient object. Second compositional attribute is the rule of thirds which states that the salient object should be placed near or at the intersection of the three horizontal and vertical lines which divide the image equally. This property is also implemented using the salient object predictor. Third compositional attribute is the low depth of field (DoF) which states that the interesting object should be in focus while the rest of the image should be out of focus. This attribute is implemented using low DoF classifier which uses an SVM trained on Daubechies wavelet based features [23]. The last composition attribute is the presence of opposing colors for which an SVM is trained based on a specifically built color histogram from 1000 training images.

#### 3.2 Content attributes

Content attributes consist of the presence of people, portrait depiction, presence of animals, indoor-outdoor classification and scene type. For the presence of people and portrait depiction, the Viola-Jones face detector [24] is used. For the presence of animals, indoor-outdoor classification and scene type attributes 17 SVM classifiers are trained using the intersection kernel computed on spatial pyramid histograms [25] (1 each for animals, and indoor-outdoor, and 15 for various scene categories).

#### 3.3 Sky illumination attributes

Finally for sky illumination attributes a given image is divided into three regions namely sky, vertical and horizontal geometric classes based on the work of [26]. SVM classifiers are trained on low level features extracted from 1000 manually labeled images to predict three type of sky attributes. These attributes are clear skies, cloudy skies and sunset skies.

To estimate the aesthetics of the images, an SVM classifier is trained where the input image representation is the outputs of the 26 high level describable attribute classifiers. The image data set used for aesthetic evaluation consists of images rated by humans with respect to aesthetics. The same method is applied to predict the interestingness of an image. The image data set used for interestingness evaluation contains images that are more socially interacted by people. The attributes based aesthetics and interestingness prediction is shown in Figure 3.

![Figure 3. Attributes based aesthetics and interestingness prediction](image)

### 4. Attributes based Image Retrieval

Image retrieval methods based on local features [27] [28] [29] extract local features from the query image and then compare with those of the database images. Images with the most number of matches are considered to be similar. However, these methods are not in accordance with the human description of objects i.e. the way humans describe objects by their appearance, shape and relationship with other objects in the images. Image retrieval methods based on understandable attributes of objects will be most satisfy-
ing for humans. In this section we present recent methods that use attributes for image retrieval.

### 4.1 Image retrieval based on classemes

Torresani et al.[30] have used attributes which they call classemes for image retrieval. Images are described by low dimensional descriptors whose components are the outputs of the category-specific classifiers applied to the image. Their approach is similar to [2] and [3], however the images they used are not annotated and their attributes have no specific semantic meanings. The images are independently obtained from a search engine for each category from the LSCOM ontology[31]. The feature vector is learnt using the LP-β kernel combiner[32] with 13 kernels for 13 types of features to represent an image and this vector is called the classemes vector. These classemes are then shown to perform well in transfer learning by retrieving 'similar' images for a novel category image.

### 4.2 Image retrieval based on compact combination of classemes and fisher vectors

The attribute descriptors of Torresani et al.[30] and Fisher vectors based on low-level image features are used by Douze et al.[33] for image retrieval. Apart from these descriptors, textual features are also extracted from the tags and text around the images. Both Fisher vectors and attribute descriptors are normalized and combined. The performance of this combined descriptor gave a performance on par with the state-of-the-art Fisher[34] and VLAD[35] descriptors with a somewhat lower dimensionality. Dimensionality reduction of both the Fisher vectors and attributes descriptor is done and evaluated as well and it is noted that relatively lower dimensions of the descriptors gave comparable results as well. Additional performance and compactness are also obtained by encoding the image descriptors resulting in better performance over the state-of-the-art methods. This approach is shown in Figure 4.

### 4.3 Explicit modeling of attributes interdependence for multi-query image retrieval

Attributes correlation is sometimes unwanted in transfer learning because an object like “wheel” can be metallic as well as wooden. So an attribute classifier for wheels trained on images of cars will not give satisfactory results if presented with the image of a wooden cart. But Siddiquie et al.[36] took the advantage of attributes correlations and used it to get a better performance. For example a person having beard is mostly likely to be a man and not a woman. They explicitly model the correlations and relationships of the attributes for image retrieval and ranking based on multi-attribute queries. Image retrieval and ranking are formulated using the same framework which also supports multi-label queries. For a multi-attribute query, the prediction function returns a subset of relevant images which maximizes the score over a weight vector. This weight vector models a) the confidence scores of all individual attributes in the query over all images and b) the correlation between the individual attributes of the query and all the attributes. The weight vector is learnt over the training images and their respective labels using a max-margin function by optimizing the training error based on a number of performance metrics. For ranking, a permutation of a set of images is returned by the prediction function for a multi-attribute query. The weight vector is again learnt from the training images along with their corresponding labels using a max-margin function by optimizing the incurred loss which is given by the normalized discount cumulative gain (NDCG). The proposed method outperformed two state-of-the-art methods on three benchmark data sets in retrieval and outperformed several methods on the same data sets in ranking.

### 4.4 Multi-query image retrieval based on weak attributes

Using an approach similar to Siddiquie et al.[36], Yu et al.[37] modeled the interdependence of the attributes in multi-attribute queries for large scale image retrieval based on ‘weak attributes’. They are called weak attributes because of the fact that they may or may not be directly related to the query attributes. A retrieval model is learnt for a multi-attribute query using max-margin training formulation on training data. This model is used to predict a subset of images for the given query. However a query

![Figure 4. Attributes based aesthetics and interestingness prediction](image-url)
adaptive selection of weak attributes is needed due to the fact that a small set of weak attributes is only related to the query attributes among the large pool of weak attributes. To model this mapping, a two-layered semi-supervised graphical model is developed. A supervised layer is constructed using the training data with query attributes labels and an unsupervised layer is constructed using the test data with no query attributes labels. The inference is done iteratively on both the layers in an alternating fashion. In this work, they have constructed the largest multi-attribute data set to date as well called the a-TRECVID data set. The method has been shown to outperform several state-of-the-art methods on several benchmark data sets.

5. An Overview of the Attribute Data Sets and Future Directions

Data sets play important role in the evaluation of solutions to computer vision problems. They usually consist of annotated images which are used for training the model and as ground truth. This section gives an overview of the attributes based data sets.

- **a-Pascal and a-Yahoo Data sets**
  This data set\(^3\) is created and used by Farhadi et al.\(^2\) for the description of objects based on attributes. It further consists of a-Pascal data set and a-Yahoo data set. The a-Pascal data set is created from PASCAL VOC 2008\(^3\). In order to classify and detect the visual objects classes, the PASCAL VOC 2008 was created from images taken under different conditions of pose, viewpoint and orientation. Around 150-1000 objects from each of the 20 categories along with 5000 instances of people are present in the data set. 64 attributes are used to describe the objects. Amazon’s Mechanical Turk is used for annotations. The a-Yahoo data set is created to supplement the a-Pascal data set. The images of 12 additional categories are collected using the Yahoo image search. They are also annotated with same set of attributes. The object categories included in a-Yahoo data set are similar to those of the a-Pascal data set. This is done for the evaluation of the predictor’s ability for cross-category generalization.

- **Animals with Attributes Data set**
  This data set\(^4\) is created in the work of Lampert et al.\(^3\) to provide a platform for the benchmark algorithms of transfer learning. It consists of 30475 images from 50 categories of animals and marked with 80 attributes defined by Osherson and Wilkie \(^3\). The images are collected from image search engines, Google, Flickr, Yahoo and Microsoft by using the animal names.

- **Cross Category Object Recognition (CORE) Data set**
  This data set\(^5\) is developed by Farhadi et al.\(^2\). The purpose of this data set is to localize and describe objects. It contains 2780 images from ImageNet data set \(^4\) with 3192 objects from 28 categories. The annotations include object segmentation, segmentation of category, part labels and parts, masks for common materials, viewpoint and pose. The annotations consist of 26695 parts of 71 types, 30046 attributes of 34 types and 1052 material images of 10 types. This data set is not as challenging as PASCAL VOC\(^3\) because the purpose of this data set is to study the problem of cross-category generalization.

- **UIUC PASCAL Sentence Data set**
  This data set\(^6\) created by Rashtchian et al.\(^4\) consists of images for which a one sentence description is obtained from annotators using the Amazon’s Mechanical Turk. The quality control of the annotations is done by recruiting only those annotators who manage to pass a qualification test.

- **SBU captioned Photo Dataset**
  This data set\(^7\) created by Ordonez et al.\(^5\) contains 1 million captioned images from flickr. These captions are filtered that they contain at least two words from the keyword list and at least one spatial preposition.

- **Attributes Discovery Data set**
  This dataset\(^8\) created by Berg et al.\(^6\) consists of 37705 annotated images from 4 shopping categories which are ties, earrings, bags and shoes. The textual description is associated with images collected from like.com.

- **PubFig: Public Figures Face Database**
  This data set\(^9\) created by Kumar et al.\(^7\) is a large face data set which consists of 58797 images of 200 people collected from various sources on internet. These images are taken in uncontrolled environment and there is a great variation of pose, illumination, expression, scene and imaging conditions. They have annotated the images with attributes selected from a list of 65 facial attributes using the Amazon Mechanical Turk.

These data sets provide annotated data for future research. Some future directions regarding solutions based on attributes are given as follows.

- **Attributes** are shown to predict the aesthetics of an image as reviewed in section.3. However it will be interesting to know the semantic attributes that makes an image memorable. A first attempt about the image

---

\(^3\)http://www.ee.columbia.edu/~dyvim/weak/ last visited Sep, 2012
\(^4\)http://vision.cs.uiuc.edu/attributes/ last visited Sep, 2012
\(^5\)http://vision.cs.uiuc.edu/CORE/ last visited Sep, 2012
\(^6\)http://vision.cs.uiuc.edu/pascal-sentences/ last visited Sep, 2012
\(^7\)http://dsl1.cewit.stonybrook.edu/ vicente/sbucaptions/ last visited Sep, 2012
\(^8\)http://tamaraberg.com/attributesDataset/index.html last visited Sep, 2012
References


[41] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting Image Annotations Using Amazon’s Mechanical Turk. In NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk., 2010, 139-147.


