Implementation Exercise

Modeling the shape of a scene : A holistic representation of the spatial envelope

Brief Summary of the Paper :

1. The paper proposes a model for recognising real world scenes using a high level of abstraction which uses only the dominant spatial characteristics of the image.
2. This approach is different from other approaches which use the local features like oject recognition/face detection. Here, we are using features which have lowest visual resemblance to the actual image.
3. Each image is given a certain value for each of the parameters (openness, naturalness, roughness and expansion) called the GIST descriptor of the image using functions described in the paper. These functions were trained using different training data.
4. Both, the descriptor for the entire image(DFT) and local descriptors(WFT) are taken by sampling the image at different spatial locations.

Training Data

The dataset used contains 8 types of scenes for classification. Total number of features for each image is 512. Each image is sampled at 4x4 subparts and features are taken for 32 orientations of each subpart giving 512 features in total.

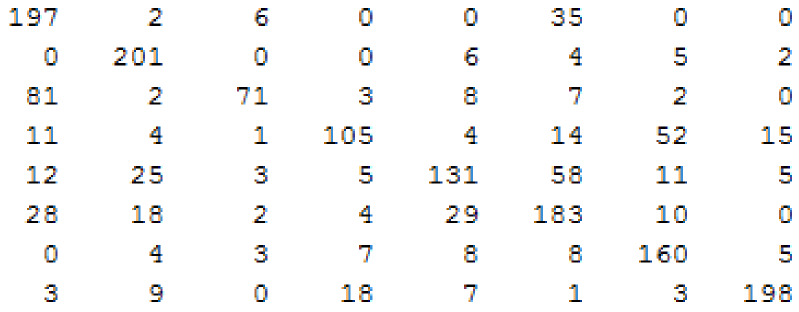
Coast	360
Forest	328
Highway	260
Inside City	308
Mountain	374
Open Country	410
Street	292
Tall Building	356
Total	**2688**

As each class is not equally represented, I took 30% of each for the training data(892 samples) and left the rest for testing (1796).

Test Results

Using SVM [Gaussian Kernel : Implemented using LibSvm] - 69.37% accuracy (1292/1796 classified correctly)

Confusion Matrix



```
   197     2     6     0     0    35     0     0
     0   201     0     0     6     4     5     2
    81     2    71     3     8     7     2     0
    11     4     1   105     4    14    52    15
    12    25     3     5   131    58    11     5
    28    18     2     4    29   183    10     0
     0     4     3     7     8     8   160     5
     3     9     0    18     7     1     3   198
```

Possible reasons for low accuracy and suggested improvements:

1. A large number (81) of the ‘highway’ class are wrongly classified as ‘coast’. This may be due to similar spatial features in each :
    a. Horizon gives a dominant pattern in the resulting fourier transform for both
    b. Spatial features for both may be similar since both may have a smooth sky background for a major part of the image.
    c. The pattern on highways and the seawater/coast may be similar (similar lines pattern) leading to similar GIST features.
2. We could use the colours in the image to get more features rather than just relying on patterns. Natural images are expected to be more greener than city images, coastal images will have more areas of blue than others and so on.

Brief Understanding and Thoughts on Given Papers

1. Zero-Shot Learning via Visual Abstraction

    - Some concepts are difficult to describe semantically to the computer. This paper proposes that abstract illustrations made using an interface to represent certain interactions can be used to classify real world images for these interactions using Zero Shot Learning.
    - It has two parts, one being an instance based learning where the training data consists of images being classified into semantically labelled classes and the other being category level learning where illustrations are used to learn classification of real world images.

- Features based on the angles of different body parts, gaze of the people involved, expressions, gender, position of the limbs etc have been used.
- To test whether automatic pose detection improves the results, a pose detector has been used (for PARSE dataset) and the results show significant improvement.
- To test if a learned mapping from the features of the abstract illustrations to the features of real world images, GRNNs were used and results show some improvement for PARSE dataset but not much for the INTERACT dataset.

Thoughts

- This method of using abstract ideas of concepts to train real world concepts could be used in **learning emotions from sound input**. For e.g. Shouting, Crying, Laughing etc.
- Zero Shot Learning gives us the advantage of learning features for classes which may not be in the training set. It allows classification of cases which may have very less chance of occurring or for which training data cannot be collected.

2. Bringing Semantics Into Focus Using Visual Abstraction

- This paper proposes that real world images are not necessary for understanding semantic concepts. The same can be learned from abstract illustrations created using clip art.
- A version of information gain is used to find out which features provide most information about the class label. Both Mutual Information and Conditional Mutual Information for pairs of features were used.
- MI and CMI scores were calculated for features like occurence of a features, attributes like expression, co-occurence of features, spatial locations and depth.
- This approach was compared with low level features such as GIST and SPM and shown to be a significant improvement (~60% accuracy over ~10% for k=10 nearest neighbours).

Thoughts

- In both of the papers (previous and this one) a novel approach that abstract illustrations can be used to learn real world concepts has been used. This is advantageous as annotation in real world images is difficult to obtain and even if it is obtained, there is still a degree of ambiguity which comes from human annotation. Also illustrations allows creation of a vast dataset for describing each concept which may not be possible in real world images.
- The findings of the paper may help finding semantically important features in real world images by comparing with the features which were found out to be important in abstract images.

3. Interactively Guiding Semi-Supervised Clustering via Attribute-based Explanations

    - Image clustering is a difficult problem as description of required clusters cannot be figured out. Hence semi supervised clustering is used to train the algorithm with the description of clusters required by the user.
    - The approaches before this paper have tried semi supervised methods by having a user indicate whether to images are similar or not for iteratively selected pairs to train the clustering algorithm. But this requires a huge training set to achieve even a small accuracy. [5000 images for 30% accuracy in one of the papers] Also these constraints are a weak indicator of the clustering.
    - This paper tries to get attribute based feedback instead from the user which describes why the images are similar or not.
    - The user is provided with images which are neither too similar neither too dissimilar(pair with highest entropy) to obtain constraints since these will be most informative for further clustering.
    - Soft constraints provide the best results for the clustering.

    Thoughts

    - This approach allows users to train the model with their desired clustering parameters which may be different for other users.
    - Though Binary attributes may not learn more information over the iterations, relative attributes help to train the model with every iteration since the user selects many different values.

4. Towards Transparent Systems: Semantic Characterization of Failure Modes

    - The paper discusses about characterising failure modes of a recognition system by semantic attributes. For eg. it may fail in low quality images, harsh lighting conditions, side view of face etc.
    - Discriminative Clustering is used to find attributes in the mistake images set.
    - Weighted Linear regression ensures that attributes present most in that cluster get more preference, thus getting more images with similar attributes clustered in that cluster.
    - Hierarchical clustering is used to form a tree type specification chart for the attributes which is easier to navigate but it may perform slightly worse than simple clustering.

Thoughts

- It would help researchers find out for which modes the model repeatedly fails so that more training data can be collected and the probability of failures is decreased.
- This approach helps in finding whether the predictions made by a machine can be trusted or not and with what accuracy. Also it helps the machine prepare for oncoming failures.
- The user can find out when to ignore a prediction from the machine and what percentage of the predictions may fail. Thus a machine which cannot be used in many cases may be very accurate when it is used whereas a machine which fails rarely may not accurately discriminate between the classes. Thus there is a accuracy vs number of uses tradeoff here.

5. Visual Attributes for Enhanced Human-Machine Communication

- This paper discusses using attributes to learn concepts and features.
- Attributes may help to teach certain concepts which humans understand through common sense like in the first paper. Currently, human provided labels are coarse i.e. they signify whether an image belongs to a class or not but does not give a description as to why it does not belong or belongs to the class.
- Attributes can easily be annotated by crowdsourcing.

Thoughts

- In all of the papers, attributes have a been a strong basis for learning the desired concept. The results are significantly better than the traditional approaches as shown in each paper. Thus attributes seem to be effective in communicating human understandable concepts easily and efficiently.
- Some errors may crop in due to the human involvement while annotating the attributes but this can be reduced largely by taking opinions from multiple users and weeding out the errors.

6. WhittleSearch: Image Search with Relative Attribute Feedback

- This paper discusses ‘Whittle Search’, a technique to iteratively search images using feedback from the user after each iteration to refine the search. The feedback consists of high level properties of desired images.
- This approach is different from other search techniques where the feedback is on which images are relevant and which are not since here we get the exact attributes which the user is looking for or is not looking for.

- A ranking function was learnt for each relative attribute to predict the extent to which it may be present in an image. This was done before the search is presented to the user.
- The feedback results in focussing on a particular part of the M dimensional attribute space.
- Finally, a hybrid method combining use of both relative and binary attributes is used.

Thoughts

- The user doesn't need to worry about the algorithm used since the feedback given is a high level description instead of parameters to tune the algorithm.