ABSTRACT
We have decided to focus on the crowdsourcing aspect of the project. We report our progress.

Categories and Subject Descriptors
I.4.9 [Image processing and Computer vision]: Application

General Terms
Management, Design, Experimentation, Human Factors

Keywords
Image recognition, crowd sourcing

1. INTRODUCTION
We are focusing on the crowdsourcing aspect of the original proposal. The website is designed to break up visitors into 4 groups:

<table>
<thead>
<tr>
<th></th>
<th>No Conflict Score</th>
<th>Conflict Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>Popular</td>
<td>Conflict Group</td>
</tr>
<tr>
<td>Popular</td>
<td>Popular</td>
<td>Both</td>
</tr>
</tbody>
</table>

If a user is in the popular image row, then they will see images weighted on their popularity; otherwise they see random images. If the user is in the conflict score column, they will see images weighted by their conflict score (number of classifiers that got the images wrong). If a user is in the “Both” group, they see images weighted jointly with both popularity and conflict (which are assumed to be independent).

2. PROGRESS
We discuss what we have accomplished in our proposed implementation plan up and also what not. In addition, we discuss issues and problems that we encountered.

2.1 Completed

2.1.1 Data Processing
Originally, we scraped ~162,000 images from Tumblr. We completed processing ~82,000 of these. Here is the breakdown of the number of images and corresponding classes for each:

Of these, we have computed hog and sift features, although there is significant data corruption.

2.1.2 Web Platform
URL: downey-n1.cs.northwestern.edu/~northanapon/colabel
We completed a web platform for crowdsourcing. The website will random images in the dataset for users to label based on our scoring criteria.

2.2 Not Completed

2.2.1 Data Processing
Clustering (ie, vector quantization) is in progress. This should take ~16 hours. We anticipate that bag of words computation to take ~9 hours.

2.2.2 Classification Result
Once clustering and bag of words computation is completed, then classification should take minutes.

2.2.3 Launching the website
The classification result will be used to compute scores based on 4 groups of scoring criteria, then we can launch the website.

2.3 Issues
- Corrupted data: Due to the distributed feature detection system we occasionally get collisions where two computers try to write to the same file at once. Corrupted files can be detected quickly using numpy’s built in load function.
- Space Limitations: We are running out of space on QUEST. We have used 1416.3GB of our 1500GB quota.
- Speed Limitations: Previously, we were using k-means to cluster data. This algorithm runs in $O(n^2 \log(n))$ time. We have switched to “bisecting k-means” instead.

3. NEXT MILESTONES

<table>
<thead>
<tr>
<th>Milestone</th>
<th>Members(s)</th>
<th>Date</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Download Images</td>
<td>Thanapon</td>
<td>November 16</td>
<td>Completed</td>
</tr>
<tr>
<td>Comparison of SIFT and HOG features using offline training</td>
<td>Thanapon, Matthew</td>
<td>November 21</td>
<td>Abandoned; but features extracted</td>
</tr>
<tr>
<td>Implementation of KD-Trees for performance enhancement</td>
<td>Matthew</td>
<td>November 21</td>
<td>Abandoned; not necessary</td>
</tr>
<tr>
<td>Web application for labeling</td>
<td>Thanapon, Matthew</td>
<td>November 23</td>
<td>Platform complete; waiting for data</td>
</tr>
<tr>
<td>Compute Results from Classification</td>
<td>Matthew</td>
<td>December 2</td>
<td>In Progress</td>
</tr>
<tr>
<td>Have crowd sourced labeled data (from friends if possible)</td>
<td>Thanapon, Matthew</td>
<td>December 7</td>
<td>Postponed</td>
</tr>
<tr>
<td>Analysis of experimental data</td>
<td>Thanapon, Matthew</td>
<td>December 7</td>
<td></td>
</tr>
</tbody>
</table>

24257
25817
5483
480
28190