# Automatically Annotating the MIR Flickr Dataset

## Experimental Protocols, Openly available Data and Semantic Spaces


**Jonathon S. Hare and Paul H. Lewis**

**School of Electronics and Computer Science, University of Southampton, UK**


## An Extended Protocol for the Evaluation of Automatic Image Annotation Using MIR Flickr

The availability of a **large, freely redistributable** set of **high- quality annotated images** is critical to allowing researchers in the area of **automatic annotation**, **generic object recognition** and **concept detection** to compare results.

The recent introduction of the **MIR Flickr** dataset allows researchers such access. A dataset by itself is not enough, and a set of **repeatable guidelines** for performing **evaluations** that are comparable is required.

It is also useful to compare the **machine-learning** components of different **automatic annotation** techniques using a **common** set of **image features**.

To this end, we have produced a **protocol for performing annotation experiments** with the MIR Flickr dataset, together with a set of **visual-term features downloadable from our website**.

### MIR Flickr: Dataset of 25000 Annotated Images

3 Training Collections:
- #1: 5000 Images
- #2: 10000 Images
- #3: 15000 Images


### Automatic Annotators

**Training and optimisation performed on each of the three training sets**

### Performance Evaluation

Evaluation of annotator performance on the test collection using **standardised procedures** and **publicly available tools**. Reporting of **computational efficiency** and **implementation details**.

#### Precision/Recall Statistics (using `trec_eval`)

**Perform a hypothetical retrieval experiment for each annotation term in turn and use `trec_eval` to produce:**
- Interpolated precision-recall graphs.
- Average precision per term.
- Graphs of precision versus number of images retrieved (up to 1000 images)

#### ROC Analysis [AUC and EER] (using `eval_tool`)

**Use the `eval_tool` tool from the ImageCLEF 2008 and 2009 visual concept detection task to perform a ROC analysis and produce:**
- Area Under Curve (AUC) values for each annotation term.
- Equal Error Rate (EER) values for each annotation term.

#### Computational Details

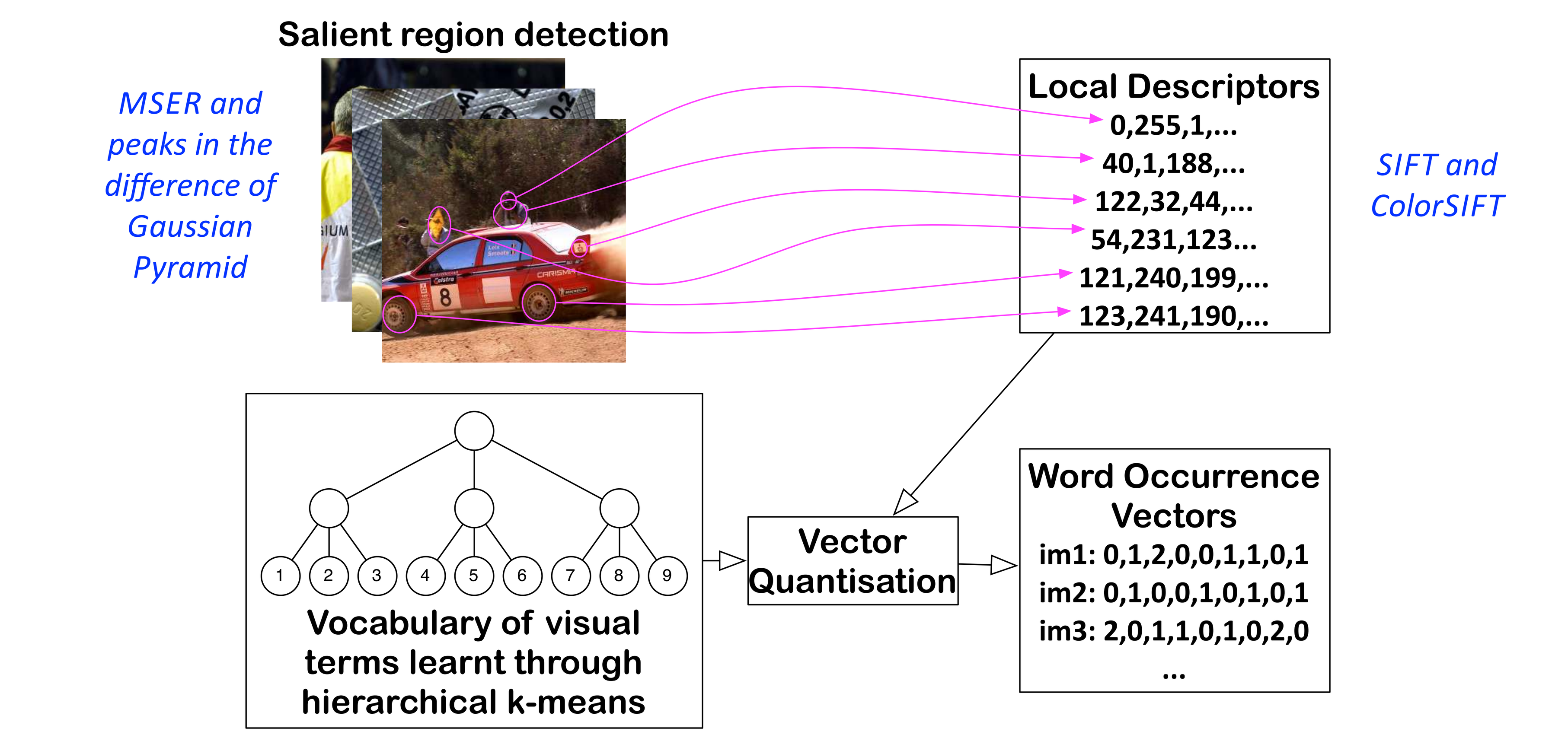
**Provide details about your implementation and computational efficiency:**
- Time taken for feature extraction and time taken to train the annotator.
- Software and languages used.
- Hardware setup (single CPU, multithreaded, cluster, ...), etc.

## Tools, pre-generated features and ground truth (for `trec_eval` and `eval_tool`) available from: http://users.ecs.soton.ac.uk/jsh2/mirflickr

## Demonstration of the Protocol using our Semantic Space Annotator

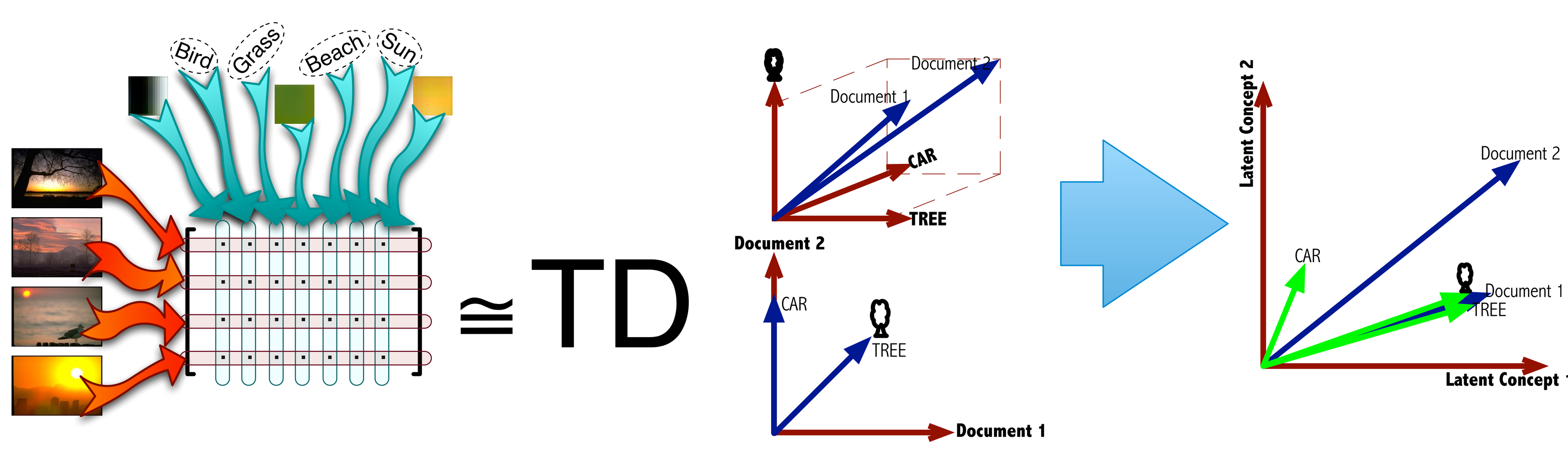
### Image Features

We used a **bag-of-visual-terms** feature morphology for the annotation task. We combined **MSER** and **difference-of-Gaussian** salient regions with **SIFT** features and **MSER** regions with **Colour-SIFT** features, using **3125** term vocabularies learnt using **hierarchical k-means** for each detector/ feature combination. The actual feature used for experimentation was a **concatenation** of the three visual-term vocabularies into **one large bag of 9375 visual terms**.



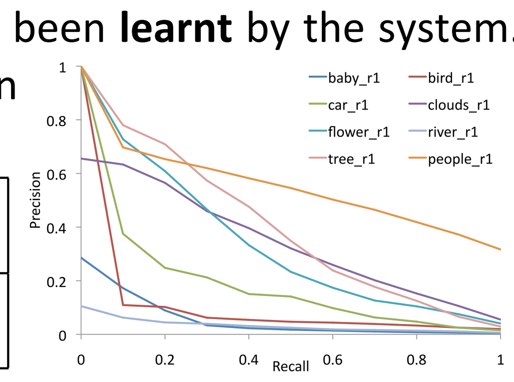

### Annotation Technique

We used an **auto-annotation** tool that we had previously developed. The tool uses a matrix factorisation of a multi-lingual (visual-terms and keywords) term-document matrix to build a **semantic space**. Un-annotated images can be projected into this space (based on their visual-terms), and their placement is such that they occur **"near"** keywords that describe their content.



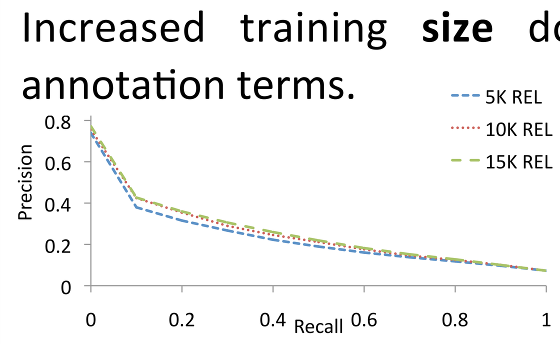

### Experimental Results

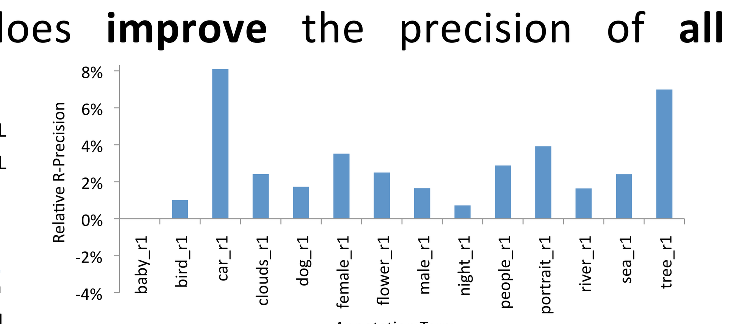
As is typical in auto-annotation experiments, there is much **variation** in how well a given **term** has been **learnt** by the system. Overall, the scores are **reasonable** given the **simplicity** of the annotator.




| Training set size | Annotation Set ALL EER | ALL AUC | POT EER | POT AUC | REL EER | REL AUC |
|---|---|---|---|---|---|---|
| 5000 | 0.319 | 0.742 | 0.331 | 0.727 | 0.296 | 0.772 |
| 10000 | 0.315 | 0.748 | 0.326 | 0.733 | 0.283 | 0.789 |
| 15000 | 0.303 | 0.761 | 0.318 | 0.743 | 0.272 | 0.797 |

Surprisingly, the **effect of training set size** is only **marginal**. Increased training **size** does **improve** the precision of **all** annotation terms.



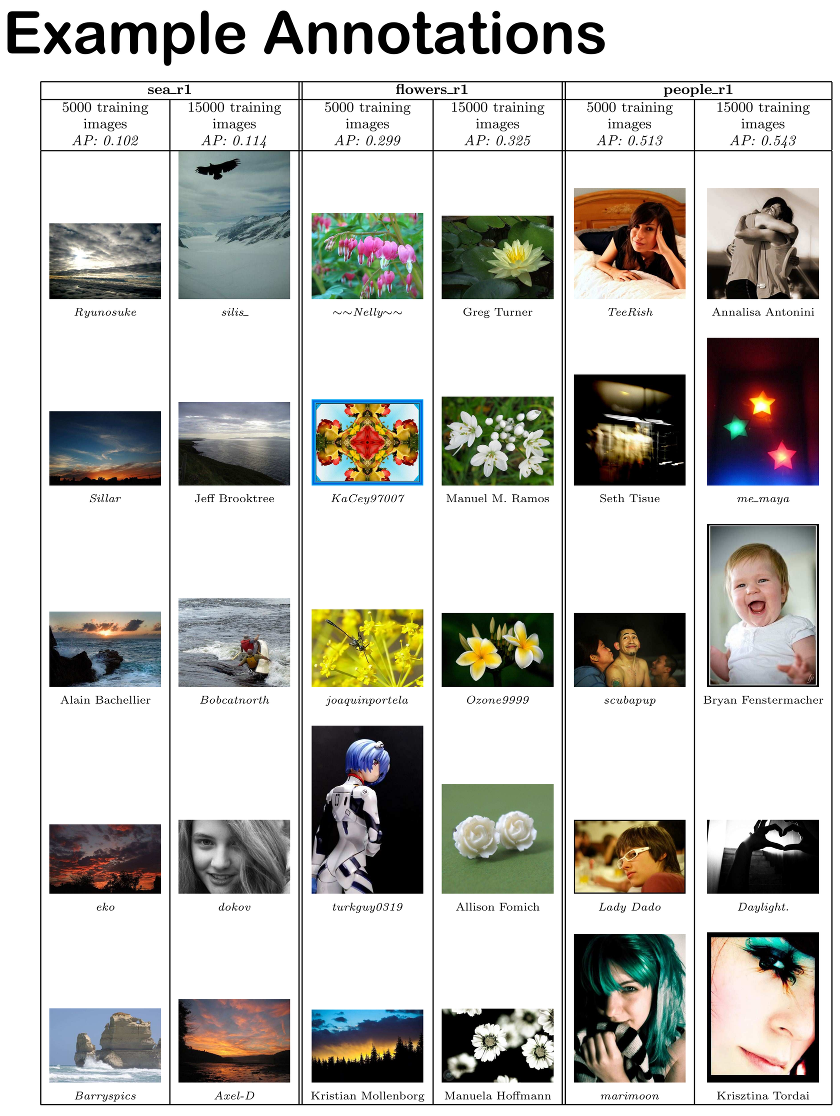




### Computational Performance

Generating features is an **embarrassingly parallel** problem and can be **easily scaled**; for an average **single** image we estimate it takes about **5.9** seconds to generate the bag-of-visual terms representation.

Our automatic annotator can be **trained** in around **5 minutes** with **5000** training images and **10 minutes** with **15000** training images.

### Example Annotations




### Summary of our findings:

- With our semantic-space annotator, **increasing the size** of the training collection does **improve annotation performance**, but only by the **smallest** of margins.
- The results from **ROC** analysis can **differ** greatly from **Precision/Recall** analysis; the two approaches to analysis measure very different things.
  - Choosing to optimise the ROC statistics (AUC/EER) when the end goal is retrieval is **not** a good idea.
  - **Optimising** for **increased precision** will inherently **increase AUC** and **decrease EER**.
- We would expect state-of-the-art annotators, such as those based on multiple (one per term) **SVMs** to perform **better** than our annotator using the **same features**.


IAM — http://iam.ecs.soton.ac.uk

living knowledge — http://livingknowledge-project.eu/

LiveMemories — Active Digital Memories of Collective Life — http://www.livememories.org

University of Southampton, School of Electronics and Computer Science — http://www.soton.ac.uk