Determining Spatial Classification Models for Automated Landmark Identification

Mark Hughes

Center for Digital Video Processing
Dublin City University

ABSTRACT

The use of interest point detection and key point descriptors has been used successfully for object recognition and for automated image matching. We aim to use keypoint descriptors fused with spatial data to automatically identify landmarks and buildings within images using a large-scale training database. However one big problem that exists in large-scale image databases is that an average image can yield around 1000 keypoints. To compare each feature from one image to every feature extracted from all the images within a large-scale database is extremely computationally expensive and requires a lot of time to execute. Therefore a technique is required which will provide reliable image matching using these extracted keypoint values in an acceptable timeframe. In this proposal we describe the use of spatial filtering fused with classification models based on interest points. We aim to cluster related images and train support vector machine classification models based on these image clusters interest point values then assign spatial locations to each of these models.

1. INTRODUCTION

The main outline of this proposal is to enhance automated image tag and caption creation using interest points and classification models. Each commonly used interest point detection method will generate on average up to 1000 keypoints within each image. This presents a considerable challenge in terms of matching two images using their image points and computational overhead. To match each keypoint from one image to each keypoint in every image located in a large-scale database is extremely computationally expensive. To put it into perspective to compare one image to all images in a 1000 image database using the sift algorithm would require 128 million comparisons to be made (1000 images * 1000 keypoints * 128 values per keypoint vector). To compare one image against a database of 100,000 images would require over 12 trillion comparisons to be made and this number would grow considerably as the size of the database grows. Clearly this type of point to point matching could not be done in real time. A new technique is thus desired which will filter the amount of keypoints that needed to be compared or a new technique, which doesn't actually match keypoint by keypoint singly in order to be able to do this matching in real time. In this proposal an approach that uses SVM classification models with assigned spatial data is described.

2. BACKGROUND

Almost all widely used image search engines today use user-defined or semi-automated image tags to retrieve and rank images which are deemed to be relevant to a users query. The accuracy of the results returned depends largely on the accuracy and richness of the tags associated with the image. The main disadvantage of this approach is that most casual users will not spend the appropriate time required to create accurate tags for images. Another disadvantage is that a lot of user created tags could be heterogeneous and might not be applicable for other people's ideas of the objects or places located within an image. A lot of research work is currently taken place to analyse images and using image content attempt to automatically create semantic tags. Using low-level image features alone has been shown to work well only with very low-level semantics such as distinguishing between outdoor and indoor images [1] and distinguishing between cityscapes or landscapes [2]. Low-level image features don't seem to distinguish accurately between high-level semantics. We discuss a new approach that will utilise image and object matching techniques using interest point detection and keypoint descriptors fused with SVM's and spatial data.

The approach that we propose is to enable efficient image matching using image classification models fused with spatial data. Multiple image views of landmarks taken from similar viewpoints will be clustered to create a single view model. With this model we plan to classify other images as belonging to the same cluster and hence create an image tag automatically for the new image. This new image could also be added to the cluster and a new model could be created to replace the old model, which would be more robust and accurate. Therefore the more images uploaded to a system based on this approach, the more accurately the system would function.

There are two main advantages to be gained from clustering multiple image views into single models:

- Computational overhead: The amount of time taken to compare and classify images in a large-scale database is drastically reduced. With efficient filtering methods this classification could be theoretically done in real time in large-scale databases.

- Robustness: Increased robustness is obtained by combining features obtained under multiple imaging conditions into a single model view.

3. PROPOSED APPROACH

A large-scale training database is to be created. This database contains images, which have been manually tagged and contain spatial data regarding the location where the image was taken. All of these images will then be clustered based on spatial data eg. all images based within a 500-meter radius of a location will be clustered together. These clusters of images will then be split into sub-clusters based on image content. Low-level image features will be fused with local image features based on the SIFT [3] and SURF [4] algorithms to create these sub-clusters. Each of these clusters is intended to represent a view of a building or landmark from a specific viewpoint. Using these image features a SVM classification model will be trained for each cluster. Each of these classification models will be assigned spatial coordinated based on a mean of the spatial data from all images located within the cluster.

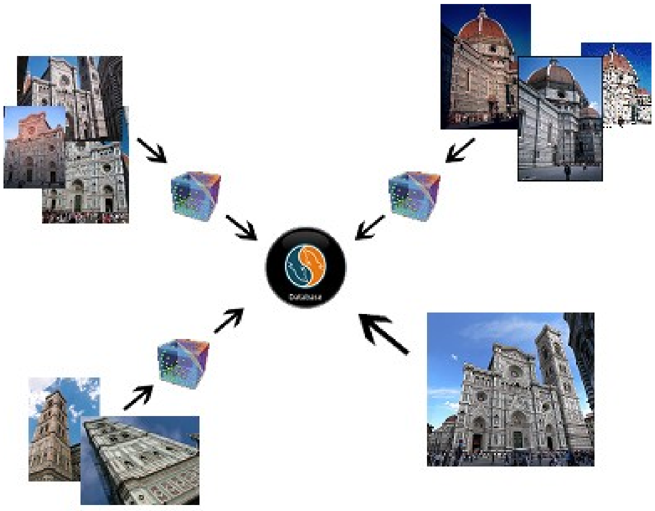

Figure 1. *Brief outline of proposed approach on a small scale. Image features extracted from image in bottom right are used as inputs to the classification models trained using image features from clusters of images in other three corners.*

To automatically identify a landmark within an image, local image features will be extracted and will be inputted into all the classification models which have spatial coordinates located near the images spatial location will be used. If one of these classifiers outputs a confidence value above a certain threshold then that image is tagged with the captions or tags associated with the classification model eg. 'View of Front of Christchurch Cathedral'. A big research challenge that exists here is how to automatically create these image views while ensuring that classification accuracy remains high. As interest points from occluding objects or of objects, which have been incorrectly added to a cluster, are used as positive examples in the training they will add a lot of noise to the classifier and significantly effect classification performance. It is important that the clustering process is as accurate as possible. Ideally images of the same landmark taken from the same viewpoint should only be included in each cluster.

Another challenge is how to train the SVM's. Using all the image points from each image will also create a noisy training set as certain interest points could be obtained from occluding objects and background clutter. Other approaches could be to combine local features with low-level image features as inputs to the SVM's or to use the means of interest point values as inputs.

4. FUTURE WORK

We aim to implement a large-scale system to test our hypothesis. We have already collected a large amount of training images (130,000+), which have human defined tags and spatial coordinates. An efficient method to cluster multiple images into 'image views' is required as the accuracy of the approach depends of how accurately the training images can be clustered. Different approaches to the number and combination of image features to use will be tried and tested to ascertain which works best for this problem.

9. REFERENCES

[1] Martin Szummer and Rosalind W. Picard, "Indoor-Outdoor Image Classification", IEEE International Workshop on Content-based Access of Image and Video Databases, in conjunction with ICCV'98, 1998.

[2] Rautiainen M, Seppänen T, Penttilä J & Peltola J Detecting semantic concepts from video using temporal gradients and audio classification. *Proc. International Conference on Image and Video Retrieval, Urbana, IL, 260 – 270, 2003.*

[3] David G. Lowe, "Object recognition from local scale-invariant features," *International Conference on Computer Vision,* Corfu, Greece (September 1999), pp. 1150-1157

[4] Herbert Bay, Tinne Tuytelaars, Luc Van Gool, "SURF: Speeded Up Robust Features", Proceedings of the ninth European Conference on Computer Vision, May 2006