Real-Time Large-Scale Visual Concept Detection with Linear Classifiers

Mats Sjöberg  Markus Koskela  Satoru Ishikawa
Jorma Laaksonen
Department of Information and Computer Science
Aalto University
FI-00076 Aalto, Finland
firstname.lastname@aalto.fi

Abstract

Many emerging application areas in video and image processing require real-time or faster visual concept detection. Examples include indexing of online user-generated video content and 24/7 archiving of TV broadcasts. The current state-of-the-art in concept detection uses bag-of-visual-words features with computationally heavy kernel-based classifiers. We argue that this approach is not feasible for real-time applications, and propose instead to use combinations of fast linear classifiers. In experiments with the large-scale TRECVID 2011 video database and 50 concepts, we compare several methods to improve the retrieval performance of standard linear classifiers. Fusing classifiers trained on different features and using multi-learn and homogeneous kernel maps achieve state-of-the-art retrieval precision, while retaining real-time performance even for large sets of concepts.

1. Introduction

Visual concept detection facilitates high-level querying of audiovisual data by organizing the database in terms of mid-level concepts such as objects, persons or events. Examples of such semantic concepts include images containing “cats”, videos depicting “people marching”, or “explosions or fire”. Statistical classifiers are trained on manually labelled data in order to predict the presence of such concepts in new images or video. This typically results in an estimate of the probability of a given concept being present.

Current state-of-the-art concept detectors rely on bag-of-visual-words (BoV) features, such as SIFT, and computationally heavy kernel-based classifiers, such as SVM. Unfortunately there are many scenarios today that require much faster classification, preferably real-time or better for a typically large number of concepts. One such scenario is indexing of online video: 48 hours of new video were uploaded to YouTube every minute in 2011. Another example is the Finnish National Audiovisual Archive, which archives all broadcasts of the major Finnish TV stations, and provides the public and researchers access to this material. Manual annotation of such archives is clearly not feasible, so there has been a growing interest in content-based retrieval methods in these kinds of application areas [5]. In such general domains, a large number of concepts is needed—at least in the order of hundreds—putting additional emphasis on fast detection methods.

There has been numerous approaches to reduce the computational complexity from the level of standard non-linear SVMs. Such approaches include using approximate SVM solvers [1], use of GPUs [11], and replacing the non-linear SVMs with linear classifiers. Using linear classifiers is particularly appealing, as both the classification and training time requirements are several orders of magnitude smaller. Linear classifiers as such cannot, however, typically match the performance of non-linear SVMs. To this end, this paper investigates different methods to improve the accuracy of linear classifiers while preserving the classification efficiency required for large-scale visual concept detection of video material. We look at fusing the classification results of a set of low-level visual features, using multiple linear classifiers for each feature, and utilizing approximate additive non-linear kernels [14]. All of these approaches increase the classification complexity, but the increase is negligible when compared to the computational requirements of standard non-linear SVMs.

We review the used non-linear SVMs and linear classifiers in Sections 2 and 3. Experiments using TRECVID data are reported in Section 4 and discussed in Section 5. The paper is concluded in Section 6.
2. Non-linear SVMs

The non-linear kernel SVMs can be considered the current state-of-the-art in visual concept detection, but are computationally expensive for real-time applications, as the query feature vector has to be compared to all support vectors of the model.

For non-linear SVMs, we use the the C-SVC classifier of the LIBSVM software library [2] extended to support additional kernels. As the default kernel for classifier of the LIBSVM software library [2] extended to support additional kernels. As the default kernel for

the model.

2.67 GHz processors, but all processing times are reported for a single CPU.

2.9

We extracted a total of eight features from the keyframes. As BoV-based features, we extracted SIFT and ColorSIFT features using the opponent color space [12], soft cluster assignment, and spatial pyramids with two different sampling strategies: the Harris-Laplace salient point detector (SIFT and ColorSIFT) and dense sampling (SIFTds and ColorSIFTds). The BoV codebooks were generated using k-means with 1000 cluster centroids. In addition, we used four non-BoV, non-histogram features based on the Census Transform (Centrist) [15], color moments, edge Fourier descriptors, and MPEG-7 Scalable Color (see

3. Linear classifiers

We trained all linear classifiers used in this paper using the LIBLINEAR [3] library with the $L^2$-regularized logistic regression solver, that we have observed to have a good performance with a fast run time in our earlier experiments.

3.1. Multi-learner approaches

The imbalance between the positive and negative classes is a typical property in visual concept detection. The semantic concepts are commonly sparse, and, on the other hand, the evaluation measures for concept detection focus on the top results returned, instead of overall classification accuracy.

One solution to this problem is to use sub-sampling in the negative class to obtain more balanced classification problems, and to combine the results of multiple such classifiers [9]. In the experiments of this paper, we decided to use 10 such classifiers, each trained with an equal number of random samples from both the negative and positive classes, and to use standard geometric mean to fuse the classification results.

3.2. Homogeneous kernel maps

There has been some recent advances on approximating additive non-linear kernels (Eq. 2) with linear methods, e.g. with piecewise linear approximations [6] or with homogeneous kernel maps [14].

The homogeneous kernel map (hkm) of order $n$ is a $(2n+1)$-dimensional linear approximation of any additive kernel, such as the intersection or $\chi^2$ kernel. Using the kernel map, we can encode a $d$-dimensional feature vector as a $(2n+1)$-dimensional vector and use a linear classifier with it to approximate the corresponding non-linear kernel. In this paper, we use the implementation of homogeneous kernel maps for the intersection and $\chi^2$ kernels available in the VLFeat library [13].

4. Experiments

To evaluate large-scale concept detection, we use the TRECVID 2011 dataset [8], which in total contains 19 200 videos from the Internet Archive (approximately 600 hours of video or 150 GB, and about 1 400 000 keyframes with our sampling method). 11 200 videos are used as the training data and the rest are used for testing. We use the 50 visual concepts evaluated as part of the official TRECVID evaluations. All processing was performed on a cluster of Intel Xeon X5650 2.67 GHz processors, but all processing times are reported for a single CPU.

We extracted a total of eight features from the keyframes. As BoV-based features, we extracted SIFT and ColorSIFT features using the opponent color space [12], soft cluster assignment, and spatial pyramids with two different sampling strategies: the Harris-Laplace salient point detector (SIFT and ColorSIFT) and dense sampling (SIFTds and ColorSIFTds). The BoV codebooks were generated using k-means with 1000 cluster centroids. In addition, we used four non-BoV, non-histogram features based on the Census Transform (Centrist) [15], color moments, edge Fourier descriptors, and MPEG-7 Scalable Color (see
Table 1. Processing times (secs) and MXIAP scores for the ColorSIFTds feature.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>50</th>
<th>500</th>
<th>MXIAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_{\text{GRBF}}(\cdot, \cdot)$</td>
<td>1.0</td>
<td>19</td>
<td>180</td>
<td>0.123</td>
</tr>
<tr>
<td>$k_{\text{INT}}(\cdot, \cdot)$</td>
<td>0.8</td>
<td>11</td>
<td>104</td>
<td>0.093</td>
</tr>
<tr>
<td>$k_{\chi^2}(\cdot, \cdot)$</td>
<td>0.9</td>
<td>16</td>
<td>150</td>
<td>0.085</td>
</tr>
<tr>
<td>linear</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.062</td>
</tr>
<tr>
<td>multi-learn</td>
<td>0.6</td>
<td>0.6</td>
<td>0.8</td>
<td>0.078</td>
</tr>
<tr>
<td>hkm-INT</td>
<td>0.6</td>
<td>0.6</td>
<td>0.7</td>
<td>0.094</td>
</tr>
<tr>
<td>hkm-$\chi^2$</td>
<td>0.6</td>
<td>0.6</td>
<td>0.7</td>
<td>0.085</td>
</tr>
</tbody>
</table>

Table 2. Processing times (secs) and MXIAP scores for fusion of eight features.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>50</th>
<th>500</th>
<th>MXIAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>linear</td>
<td>4.7</td>
<td>4.7</td>
<td>4.8</td>
<td>0.107</td>
</tr>
<tr>
<td>multi-learn</td>
<td>4.7</td>
<td>4.8</td>
<td>5.4</td>
<td>0.122</td>
</tr>
<tr>
<td>hkm-INT</td>
<td>4.7</td>
<td>4.7</td>
<td>5.2</td>
<td>0.126</td>
</tr>
<tr>
<td>hkm-$\chi^2$</td>
<td>4.7</td>
<td>4.7</td>
<td>5.2</td>
<td>0.125</td>
</tr>
</tbody>
</table>

our TRECVID notebook paper [10] for details). On average, the time required for extracting all eight features is 4.7 seconds per image, with 0.6 seconds for the best-performing ColorSIFTds. It should be emphasized that while the computational requirements of feature extraction are also an important aspect in real-time applications, the features used in this study were not optimized for extraction time.

We trained all applicable classifiers listed in Sections 2 and 3 for each feature. The classifiers contain only the penalty parameter $C$, except the GRBF kernel SVM that has two ($C$ and $\gamma$). For single parameter selection we use an approximate 10-fold cross-validation search that consists of a uniform sampling of the parameter space followed by another more detailed sampling in the best region. The GRBF kernel parameters were selected similarly but starting with a heuristic line search to identify a promising parameter region, followed by a grid search in that region.

The BoV features with non-linear SVM classifiers can be considered as the current state-of-the-art in visual concept detection, but with high computational requirements. Table 1 shows the processing times in seconds extrapolated for detecting 1, 50 or 500 concepts for one image with the best-performing ColorSIFTds feature and with different classifiers. MXIAP refers to mean extended inferred average precision [16]—the standard evaluation measure of TRECVID 2011—over the 50 evaluated concepts. For brevity, we omit the details here, but the other BoV features show very similar results with somewhat lower performance overall.

In Table 1, we can observe the relatively high accuracy of the GRBF kernel, with a notable difference also to the additive intersection (INT) and $\chi^2$ kernels. However, the computational costs of all kernel SVMs for large-scale concept detection are clearly unsuitable for real-time applications. On the other hand, the linear classifiers all have low computational costs. In fact, for linear classifiers the feature detection (0.6 seconds for ColorSIFTds) is dominating the processing times, even with 500 concepts. Both multi-learn and the kernel maps (hkm-INT and hkm-$\chi^2$) bring a notable performance increase over the standard linear classifiers, and the kernel map approximations reach the performance level of the corresponding non-linear kernels. Finally, intersection outperforms the $\chi^2$ kernel with ColorSIFTds and with the other BoV features as well.

Based on the 50 concept-wise results evaluated, we analyzed the statistical significance of the MXIAP results in Table 1 using paired $t$-test with $p = 0.01$. The analysis shows that the obtained differences are indeed statistically significant, except between the additive non-linear kernels and their respective homogeneous kernel map counterparts.

The performance of the linear classifiers can be further improved by using multiple features. Table 2 shows the results obtained by fusing classifier outputs of all eight features used in the experiments. The feature-wise results are fused using geometric mean. Each result has been obtained by using the type of linear classifier in question for each feature, with the exception of the kernel maps, as they are not valid for the non-histogram features, and the multi-learn results are used as replacements. The increased feature extraction required is clearly visible in the processing times for one concept, but its relative importance is reduced when we consider large concept sets.

Statistical analysis of the results in Table 2 shows that the performance of multi-learn, hkm-$\chi^2$ and hkm-INT fusion are all statistically better than that of the fusion of standard linear classifiers, and that the differences between them and ColorSIFTds with GRBF kernel (Table 1) are not statistically significant. Hence, we observe that the performance level of the GRBF-kernel ColorSIFTds can be reached with a fusion of linear classifiers, with only a fraction of the computational cost.

5. Discussion

In recent studies it has been observed that, despite their limited accuracy, semantic concept detectors can be highly useful in supporting high-level indexing and
querying on multimedia data [4]. This is mainly because such detectors can be trained off-line with computationally more demanding supervised learning algorithms and with considerably more positive and negative training examples than what are typically available at query time. Real-world multimedia retrieval requires, however, large concept ontologies to support a sufficient variety of queries. For example, LSCOM [7] currently has 834 concepts defined, and in TRECVID 2011, there were a total of 500 visual concepts defined, although only 50 were evaluated.

The ColorSIFTds feature with kernel SVMs can arguably be considered as the current standard benchmark for visual concept detection. To achieve top results in evaluations such as TRECVID 2011, it has to be further extended and fine-tuned, and combined with other features, but these sophisticated methods are typically computationally very intensive, and therefore clearly unsuitable for real-time applications.

The central issue in this paper has been to enable real-time large-scale concept detection for large amounts of image or video material. For video, a reasonable rate of analysis could be one frame per second, and from Table 2 we can see that this rate can be achieved with a set of linear classifiers for several hundred concepts while maintaining competitive detection accuracy by using about five parallel processor cores. Due to the fact that the feature extraction methods used in these experiments were not optimized for speed, most of the processing time is used in feature extraction, and further speed-up can undoubtedly be obtained by reconsidering the feature extraction components. In particular, BoV feature extraction using GPUs has been shown to be very efficient [11].

6. Conclusions

Our experiments demonstrated the feasibility of using fast linear classifiers for real-time large-scale visual concept detection. The accuracy of the linear classifiers was increased either by using multiple classifiers for each feature or with homogeneous kernel maps. In particular, the kernel map approximations achieved the same performance as their non-linear counterparts with only a fraction of the detection time.

By using fusion over several low-level visual features one can achieve the same performance as the state-of-the-art approach (ColorSIFTds with GRBF kernel). This increases processing time, due to the added features, but is still feasible for real-time performance as it remains orders of magnitude faster than the conventional approach.

Acknowledgments

This work has been funded by the grants 255745 and 251170 of the Academy of Finland, TFMC 11863 of EIT ICT Labs, and Next Media and D21 SHOK projects.

References