ABSTRACT

Local features are the building blocks of many visual systems, and local point detector is usually the first component for local feature extraction. Existing local point detector are designed with target for matching and it may not perform well when applied in image content representation. Actually many existing studies demonstrate that the simple dense sampling strategy can achieve better performance than many local point detection methods in image classification tasks. In this paper, we propose a novel point detector named semantic point detector, which detects a set of semantically meaningful patches from each image and yields more compact and complete image representation. It is learned from an set of images with concepts from a large ontology. We conduct extensive experiments based on the proposed detector, and the experimental results demonstrate the effectiveness of our approach.

Categories and Subject Descriptors
H.3.1 [Information Storage and Retrieval]: Content Analysis and Index

General Terms
Algorithms, Experimentation

Keywords
Semantic point detector

1. INTRODUCTION

Local features not only help to find correspondences in spite of large changes in viewing conditions, occlusions, and background clutter, but also yield an interesting description of the image content for image retrieval and object or scene recognition [13]. Generally, there are two main steps for extracting local features, the first is point detection, and the next is point descriptor extraction. Point detection can be viewed as a task of selecting a set of points from all possible candidates on the image grid. The widely used point detectors can be divided into two categories, namely interest point detector and dense sampling. Interest point detector is originally designed for recognizing specific objects or finding correspondence between two or more wide baseline views. It selects a set of points based on the local image structures, such as corners [4] and blobs [7], these points tend to be sparse and stable in images of the same object or scene. Recently, these detectors have been successfully used in near-duplicate detection and landmark retrieval [12, 15]. However, these detectors are designed with the purpose of finding points that are invariant under a certain family of transformations. This principle makes robust the correspondence establishment between images with the same object or scene, but it is not necessarily optimal for describing the image content and revealing semantic meanings. Actually, many recent studies have shown that the simplest dense sampling method, which provides a more complete image representation, is able to perform much better than interest point detector in many image classification tasks [10]. However, the good performance is actually achieved based on the good discriminative ability of the subsequent classifiers, as they will need to mine the useful information. Intuitively, it is not necessary to use all the patches and reasonable point selection approach is required.

In this work, we propose a learning-based point detector called semantic point detector which aims to select a set of points that can better represent the image. The point detector is built based on a patch-level binary classifier that distinguishes the informative patches from the non-informative ones. The definition of informativeness depends on the specific application. Existing local point detectors define informativeness for matching purpose. The points that are the invariant to a group of transformations are defined as informativeness, and thus different hand-designed classifiers are proposed to detect such points. Semantic point detector is designed for image interpretation tasks, and the informativeness is thus defined as the semantic representativeness.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

\textit{MM'11}, November 28-December 1, 2011, Scottsdale, Arizona, USA.
Copyright 2011 ACM 978-1-4503-0616-4/11/11 ...$10.00.
The semantic representative patches are with clear semantic meanings independent of the context. To find such patches, we learn the semantic point detector from an image dataset with the concepts from a large ontology. First, each image is represented as a linear combination of all its patches’ representations. Second, the image-level linear classifier is learned for each concept. Since these two steps only perform linear operations on the patch representation, the patch-level classifier can be directly derived from the image-level classifier. It classifies each patch according to its relevance to the concept. With a large set of such patch-level classifiers, we can distinguish a lot semantically meaningful regions in the patch space. Considering the validated hypothesis “useful patches are likely to be shared across categories” [1, 14], a large set of such patch level classifiers can form a reasonably complete semantic point detector. It selects a patch based on its semantic representativeness, which can generate a more compact and complete image representation.

As a summary, the main contributions of the paper include:
(1) we propose a semantic point detector;
(2) we suggest a semantically meaningful patch gathering process; and
(3) we evaluate and compare a number of different local point detectors on image classification and demonstrate that the semantic point detector can generate more compact and complete image representation.

The rest of the paper is organized as follows. In Section 2, we detail the proposed semantic point detector. The experimental validations are presented in Section 3. Finally, we conclude the paper in Section 4.

2. SEMANTIC POINT DETECTOR

As introduced above, semantic point detector is a semantic classifier for image patches. It is intuitive to train such classifier using a set of semantically meaningful and meaningless patches. However, such labeling process is of high cost and hard to label enough patches to well cover the semantically meaningful regions in the whole patch space. Here, we directly derive such patch-level classifiers from a set of image-level classifiers which can be easily obtained from a set of images with labels. In this section, we detail the construction process of the semantic point detector. First, we introduce the patch and image representation. Then we introduce the method to derive the patch-level classifiers.

2.1 Patch and Image Representation

Bag-of-visual-words representation has been successfully used in image interpretation tasks. However, it is with limited representation power of the patch space as all the patches belong to the same visual words are treated the same after quantization step, and many methods have been proposed to overcome the quantization problem [5, 11, 16]. Here we use the Super-Vector coding introduced by Zhou et al. [16] for patch and image representation, which achieved state-of-the-art performance on the standard datasets. This representation was shown to extend bag-of-visual-words: it is not limited to the number of occurrence of each visual word but it also provides discrimination to the descriptors belong the same visual word, which yields a better representation of the patch space.

Let $I = \{I_p\}_{p=1}^P$ be the set of $P$ local patch descriptors extract from an image, and visual vocabulary $V = \{v_k\}_{k=1}^K$ is a set of visual words generated by a unsupervised learning algorithm. The Super-Vector coding of descriptor $I_p$ is defined by

$$
\phi(I_p) = [\phi_1(I_p), \ldots, \phi_K(I_p)]
$$

$$
\phi_k(I_p) = [\gamma_k(I_p), s\gamma_k(I_p)(I_p - v_k)]
$$

where

$$
\gamma_k = \begin{cases} 
1, & \text{if } k = \arg \min_j \| I_p - v_j \| \\
0, & \text{otherwise}
\end{cases}
$$

and $s$ is a nonnegative constant and $I_p - v_k$ is the quantization error. The Super-Vector coding degenerates into bag-of-visual-words representation when $s = 0$.

Finally, image $I$ is represented by linearly combining the patch descriptors,

$$
\Phi(I) = \frac{1}{P} \sum_{p=1}^P \phi(I_p)
$$

2.2 Image and Patch Level Classifiers

For a category with a set of positive and negative images, linear SVM that employs a hinge loss is used to learn the image-level classification function $f(I) = w^T\Phi(I)$. As only linear operation are performed on the patch descriptors, by a simple mathematics transformation, the image-level classification function can be directly transferred to the patch level,

$$
f(I) = w^T\Phi(I) = \frac{1}{P} \sum_{p=1}^P w^T\phi(I_p) = \frac{1}{P} \sum_{p=1}^P f(I_p)
$$

Then we obtain the patch-level classifier, i.e.,

$$
f(I_p) = w^T\phi(I_p)
$$
Table 1: Comparison of different detectors on PASCAL VOC 2007 test set.

<table>
<thead>
<tr>
<th>AP(%)</th>
<th>DoG</th>
<th>KB</th>
<th>HarLap</th>
<th>Random</th>
<th>SPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>aeroplane</td>
<td>0.488</td>
<td>0.216</td>
<td>0.533</td>
<td>0.546</td>
<td>0.599</td>
</tr>
<tr>
<td>bicycle</td>
<td>0.277</td>
<td>0.248</td>
<td>0.421</td>
<td>0.374</td>
<td>0.475</td>
</tr>
<tr>
<td>bird</td>
<td>0.230</td>
<td>0.146</td>
<td>0.245</td>
<td>0.220</td>
<td>0.291</td>
</tr>
<tr>
<td>boat</td>
<td>0.290</td>
<td>0.306</td>
<td>0.446</td>
<td>0.474</td>
<td>0.516</td>
</tr>
<tr>
<td>bottle</td>
<td>0.114</td>
<td>0.122</td>
<td>0.095</td>
<td>0.121</td>
<td>0.124</td>
</tr>
<tr>
<td>bus</td>
<td>0.177</td>
<td>0.283</td>
<td>0.302</td>
<td>0.320</td>
<td>0.381</td>
</tr>
<tr>
<td>car</td>
<td>0.420</td>
<td>0.391</td>
<td>0.663</td>
<td>0.599</td>
<td>0.642</td>
</tr>
<tr>
<td>cat</td>
<td>0.275</td>
<td>0.274</td>
<td>0.248</td>
<td>0.318</td>
<td>0.371</td>
</tr>
<tr>
<td>chair</td>
<td>0.283</td>
<td>0.309</td>
<td>0.355</td>
<td>0.337</td>
<td>0.351</td>
</tr>
<tr>
<td>cow</td>
<td>0.139</td>
<td>0.086</td>
<td>0.223</td>
<td>0.200</td>
<td>0.266</td>
</tr>
<tr>
<td>diningtable</td>
<td>0.154</td>
<td>0.195</td>
<td>0.235</td>
<td>0.198</td>
<td>0.143</td>
</tr>
<tr>
<td>dog</td>
<td>0.212</td>
<td>0.257</td>
<td>0.210</td>
<td>0.304</td>
<td>0.322</td>
</tr>
<tr>
<td>horse</td>
<td>0.430</td>
<td>0.428</td>
<td>0.593</td>
<td>0.541</td>
<td>0.516</td>
</tr>
<tr>
<td>motorbike</td>
<td>0.219</td>
<td>0.191</td>
<td>0.457</td>
<td>0.433</td>
<td>0.430</td>
</tr>
<tr>
<td>person</td>
<td>0.637</td>
<td>0.636</td>
<td>0.699</td>
<td>0.682</td>
<td>0.717</td>
</tr>
<tr>
<td>pottedplant</td>
<td>0.126</td>
<td>0.084</td>
<td>0.114</td>
<td>0.173</td>
<td>0.113</td>
</tr>
<tr>
<td>sheep</td>
<td>0.144</td>
<td>0.148</td>
<td>0.205</td>
<td>0.235</td>
<td>0.221</td>
</tr>
<tr>
<td>sofa</td>
<td>0.192</td>
<td>0.140</td>
<td>0.202</td>
<td>0.200</td>
<td>0.281</td>
</tr>
<tr>
<td>train</td>
<td>0.346</td>
<td>0.442</td>
<td>0.492</td>
<td>0.603</td>
<td>0.480</td>
</tr>
<tr>
<td>tvmonitor</td>
<td>0.422</td>
<td>0.225</td>
<td>0.292</td>
<td>0.287</td>
<td>0.321</td>
</tr>
<tr>
<td>average</td>
<td>0.271</td>
<td>0.255</td>
<td>0.351</td>
<td>0.347</td>
<td>0.375</td>
</tr>
</tbody>
</table>

The patch-level classifier equips us with the ability of telling the semantic representativeness of a patch for the category.

2.3 Semantic Point Detector

It is intuitive that learning more patch classifiers will find more semantic regions in the patch space. Suppose we have \( C \) categories and a set of training images for each category \( c \in \{1, \ldots, C\} \), which is a typical multi-class problem and we train a set of one-versus-rest image-level linear classifiers denoted by \( \{w_c\}_{c=1}^C \), and a set of patch-level classifiers is derived via Eq. 6. The classifier \( w_c \) is learned to identify the useful patches for representing category \( c \) and these patches are denoted by,

\[
P_c = \{I_p : w_c^T \phi(I_p) > t\}
\]  (7)

then all the useful patches jointly identified by all these patch classifiers is denoted by,

\[
P = \bigcup_{c=1}^C P_c = \bigcup_{c=1}^C \{I_p : w_c^T \phi(I_p) > t\}
\]  (8)

which can cover a larger and larger semantically meaningful area in the image space as more categories are used. We define \( \max_c w_c^T \phi(I_p) \) as the semantic representativeness score of a patch, the score indicate the confidence of a patch belongs to the useful patch set \( P \).

The semantic point detector selects points according to the semantic representativeness score.

3. EXPERIMENTAL RESULTS

We first introduce the experiment for learning semantic point detector. Then we apply it to image classification. In all the following experiments, 128-dimension SIFT vectors are extracted to represent the image patches.

3.1 Semantic Point Detector Learning

We relied on the ImageNet50K as the training database for semantic point detector, which consists of 50,000 images from a large ontology of 1000 visual concepts, including about 500 kinds of living things and the other about 500 kinds of artifacts [2]. For each image, a set of 32x32 patches over a grid with spacing of 4 pixels are extracted, and a visual vocabulary \( V \) of size 1024 is trained on one million randomly sampled patches. Then each image is represented with Eq. 1, and a one-versus-rest linear SVM is trained for each category. In Fig. 3 we visualize the learned semantic representative patches of some exemplar categories from fruit and artifact.

3.2 Image Classification

To evaluate the performance of the proposed semantic point detector in applications, an image classification experiment was carried out on the PASCAL VOC2007 dataset [3]. The images in the dataset contain objects from 20 object categories. It contains 9963 images which are split into three subsets: training data (2501 images), validation data (2510 images), and test data (4952 images). All the following experiment results are obtained on the test data. We use the PASCAL toolkit to evaluate the classification accuracy, measured by Average Precision based on the precision/recall curve. Our goal is to compare different point detectors with respect to their recognition performance. To minimize other affects, we use a basic scheme to perform the classification task (not using any spatial information, not using multiple kernels, etc). The method consists of the following steps: a) a set of local points with descriptors are detected for each image; b) a visual vocabulary is constructed by clustering local descriptors; c) image representation is computed via Eq. 1; and d) an one-versus-rest linear SVM is trained for each category.

We conduct a set image classification experiments with the following local point detectors: DoG [8], Kadir Brady (KB) [6], Harris-Laplace (HarLap) [9], Random: 32x32 patches are selected randomly over the image grid, Semantic Point Detector (SPD); this is our proposed detector. Figure 2 illustrates an example of points detected by different detectors.

The visual vocabulary of size 1024 was trained on one million randomly sampled descriptors. Then each image is represented with Eq. 1, and a one-versus-rest linear SVM is trained for each category.

We conduct a set image classification experiments with the following local point detectors: DoG [8], Kadir Brady (KB) [6], Harris-Laplace (HarLap) [9], Random: 32x32 patches are selected randomly over the image grid, Semantic Point Detector (SPD); this is our proposed detector. Figure 2 illustrates an example of points detected by different detectors.

The visual vocabulary of size 1024 was trained on one million randomly sampled descriptors. Then each image is represented with Eq. 1, and a one-versus-rest linear SVM is trained for each category.
Figure 2: Examples of comparison between different detectors, the results of existing detectors (DoG, KB, and HarLap) are shown in the top row, the bottom row shows results of SPD compared with the above detectors by keeping the same number of points.

Figure 4: Mean Average Precision as a function of the number of sampled patches used for classification.

The comparison results are shown in Table 1, from which it is clear that SPD significantly outperforms the other detectors on classification accuracy.

As dense sampling has been proved to be the most effective sampling strategy for image classification. The Random and SPD will converge to dense sampling as more points sampled. We conduct another group of experiments to investigate the converge rate of the two sampling strategy. Figure 4 shows the effects on mean average precision of varying the number of patches sampled per image. It can be seen that the SPD converges much faster than random sampling, this is also demonstrate that the SPD can select a set patches that are more compact and complete.

4. CONCLUSIONS

In this paper, we have proposed a novel semantic point detector to find semantically representative patches. It is learned from an image dataset for a large ontology of visual concepts. And the extensive experiments validate the effectiveness of the new detector. Compared with existing detectors that designed for matching purpose, the semantic point detector provides a more compact and complete image representation and works much better in image interpretation tasks.

5. REFERENCES