Object Categorization based on RGBD data

Adam Kosiorek

September 30, 2013

Abstract

Abstrakt

1 Introduction

introduction

2 Bag of Words image representation

As portreyed in [14] the Bag of Words or BoW model has originated from the text retrieval domain. Originally the model enabled a vector-like representation of a text document. One of the most simple cases would be to, given a dictionary, construct a histogram depicting the incidence of words in a particular document. Such an approach obliterates any grammatical dependencies in order to retain only statistical information associated with each word. It is believed that the words' frequencies are connceted to the semantic meaning of the document. The following questions emerge:

- How does this apply to the area of computer vision?
- How can one convert an image to a text document?

In order to answer the first question one has to consider how computers process data. Digital images are nothing but streams of binary code. A simple conversion to an RGB format makes the data structured and is enough for people to understand pictures' content. Unforunately, such image representation

is extrmelly vulnerable to translation and rotation as well as changing lighting conditions. One way to address these disadvantages would be to compute locations of some characteristic points (keypoints) and describe them somehow. Even then, however, the resulting represenation is a low level one and still incomprehensible for humans. Defining robot's behaviour basing on such data might prove cumbersome, for there are too many details to consider. All of the above is reffered to as a semantic gap. It is defined by Tsai [14], with respect to the Content Based Image Retrieval or *CBIR*, as *"the gap between the extracted and indexed low-level features by computers and the high-level concepts (or semantics) of user's queries"*. Suppose a text document can be created from an image. If this is the case, then the document can be converted into a BoW model. Since the latter contains information about the semantic significance of the image an impact of the semantic gap can be reduced.

As for the second question, there is a quite well established pipeline that enables creation of text documents from images. The steps of the BoW methodology are as follows:

- Keypoint detection — keypoints are local interest points or regions. They are usually computed in such a way so as to provide scale and location invariance. Rigid transformation and illumination invariance would be desired but it is somewhat harder to achieve

- Keypoint description — each keypoint have to be described in a manner that distinguishes the particular keypoint in some way.

- Vector quantization — clustering algorithms are used to find regions in the high dimensional space of keypoint descriptors. When clusters are found, each described keypoint can be assigned to a corresponding cluster. Then, an image can be represented by the numbers of clusters, to which the image's keypoints were assigned. The numbers themselves are called *'visual words'*.

3 BoW in Computer Vision

The Bag of Words image representation has been extensively used in the areas of scene [1, 2, 14] and object categorization [16] as well as CBIR [5, 13] yielding

state-of-the-art results. Advantages of this model are simplicity, computational efficiency and at least partial invariance to affine transformation, occlusion and lighting conditions.

The Bag of Words is a type of intermediate representation. Therefore it is not sufficient to compute a BoW model of an image in order to predict its category. Additional operations are essential if the model is to be used in one of the enumerated fields.

3.1 Content Based Image Retrieval

One of the most notorious use-case of *CBIR* is to search a database in order to find an object fulfilling certain conditions. As [13] outlines, several criteria have to be met for the task to be performed efficiently. Firstly, all entries should be indexed in a concise way. Secondly, some (dis)similarity measure should be provided. Finally, an efficient search algorithm should be available.

There are numerous methods suitable for computation of objetcs' signatures. Many of them can be used in the indexing step. All the methods were divided into three general categories in [13], specifically: feature based methods, graph based methods and other methods.

Feature based methods can be either global or local. The former takes the form of a single vector or a point in a d dimensional space — the similarity measure being a point-wise distance in that space. The latter gives multiple such points for each object, rendering the computation of a similarity measure slightly more complicated. The global features takes forms of the models' volume, their mass or mass distributions. Others might incorporate the global features' distribtution — one conceivable approach would be to compute global features' distributions and summarize them into a histogram. These features have the advantage of being easy to compute and straightforward to implement. However, they are insensitive to any local shape variations, thus being ill-suited for detailed comparisions. On the other hand, they might be exploited in the preprocessing of the data with more sophisticated methods being used afterwards. Partial matching is not possible, since the global features do not encode any relations between parts of the objects.

As for the local features, [13] discusses only features describing neighbourhood of the points on boundaries of objects. Any settings that does not match

this criteria are neglected. Moreover, the authors state that the local feature based methods are inefficient and indexing is rather complex. A Bag of Words based approach, described in [5] has no such drawbacks.

Bag of Words techniques can be regarded as feature distributions, even though they are local feature based — the local features being visual words. When all the image's visual words are summarized into a histogram a distribution is created. Being similar to a point in a multi-dimensional space, it is similar to a global feature. Consequently, similar methods apply, with the distinction being that the histogram is rather a vector then a point. Thus metrics well-suited to vector comparision, such as a cosine distance, can be used.

3.2 Scene Categorization

One of the first works employing the BoW for the purpouse of scene categorization is [1]. The authors suggested a general framework. What is more, algorthims performing each main pipeline's step were proposed and evaluated as well as a codebook construction method was developed. The *Harris Affine Detector* was used for feature extraction and the *SIFT* for the feature descripton step. The visual vocabulary has been generated by clustering only a limited number of keypoints from each category. After summarizing every image in the traning set with a histogram of visual words, supervised learning was used to train two classifiers: Naìve Bayes and a Support Vector Machine. As expected the Naìve Bayes resulted in a lower accuracy than SVM (72 vs. 85% on 7 categories). An analysis on the number of centroids in the clustering step depicts that the greater the size of the visual vocabulary the greater the accuracy.

Li *et al* further refined the above approach by examining several keypoint detectors and descriptors. The main contribution of their work is, however, the development of a genuine classification algorithm based on a probabilistic graphical model. Accuracy of 76% on a large 13 category dataset was achieved.

3.3 Object Categorization

In many cases object categorization might be addressed as a scene categorization problem. There are following differences: (1) an object should be localized

on an image (i.e. its bounding box has to be found) and (2) in object categorization task information about a scene type might be used. The second case is symmetric, for in the scene categorization problem information about objects present in the scene can be utilized as well. Having said that, it is possible to consider an object categorization problem where images of singled-out objects are provided — e.g. a single object covers the majority of an image's area. Such an approach was exploited in a recent work by Zhang *et al* [16].

4 BoW based classification

The Bag of Words intermediate image represenation provides a concise way of summarizing images. It is invariant to affaine transformations, partial occlusion and, in some extent, to changing lighting conditions. What is more, it is easy to implement and computationally efficient. The efficiency can be enhanced even further with gpu based implementations as shown in [15]. In order to incorporate BoW into an object classification framework one has to feed the resulting histograms into a classification algorithms. Either generative or discriminative methods can be used. Below the main steps of BoW classification processing pipeline will be disscussed.

4.1 Detection

Characteristic point detection is the first step in any Bag of Words framework. Numerous detection methods have been developed, but choosing the right one for the particular case might prove tricky. A good overview of various mechanisms is available in [14]. The most common detectors make use of a Harris corner detector or image's first or second derivatives. A technique taking advantage of the Harris corner detector is for example a Harris-Laplace detector — the Harris function is scale adapted and its outcome is a subject to a Laplacian-of-Gaussian operator, which selects relevant points in the scale space. Images' regions' 2^{nd} derviatives — namely the regions' Hessians — can be combined with a LoG operator. This combination allows selection of points significant in the two spaces: the scale space and the Hessian's determinant space. The latter entails the speed at which pixel intensities change in the neighbourhood of a point.

A number of more complicated recipes for salient region localization have been developed and implemented. These include Scale Invariant Feature Transform [6], SUSAN [12] and Intrinistic Shape Siganutes [17]. The majority of keypoint detection formulas is being develoepd for the 2D domain. A number of them have been adapted to 3D, however. A comparative evaluation of detection algorithms available in PCL can be find in [3]. Another comprehensive study is [11].

All these formulas, called sparse feature detectors, resort to selection of maxima in specific state spaces. An entierly different scheme is to use a dense feature detector. This particular form requires users to specify a uniformly sampled grid from which points are taken. Dense detectors have an advantage of taking points from slow changing regions. A sparse detector might be unable to summarize a slow changing region such as clouds, sky or ocean. Li *et al* showed that dense detectors outperforms the sparse ones.

4.2 Description

Computing localization of a salient point is not enough. If a keypoint is to be affaine transform invariant, it has to be described in more general terms. Such description, usually in a form of coordinates in a multi-dimensional space, is provided by specialized algorithms. 128-dimensional SIFT [7] is a 3D histogram of gradient locations and orientetions. It is the most often extracted as well as one of the most effective descriptors [14]. Other methods include various color descriptors, binary descriptors such as 512-dimensional GIST [8]. There are techniques designed for 3D exclusively. Among them one can find Persistent Point Feature Histogram [10] and its faster alternative Fast Point Feature Histogram [9], both implemented in PCL.

4.3 Vector Quantization

The final step of extracting Bag of Words features is vector quantization. Generally clustering with the kMeans algorithm is used [14]. The kMeans algorithm was developed in the 50's and a variety of modifications have been developed since [4]. Some of them are: faster than the original *approximate kmeans*, *hierarchical kmeans*, which automatically chooses the resulting number of clusters and a *soft kmeans* — a variation of the algorithm that allows a fuzzy alignment

(*i.e.* each point can belong to several clusters with different weights. The soft kMeans is a compromise between kMeans' (relative to GMM, discussed below) low computational cost and GMM's precision. The number of resulting visual words depends on the number of clusters.

In order to improve performance multitude of pre- or postprocessing techniques can be resorted to. A weighting scheme such as Term Frequency (TF) or Term Frequency - Inverse Document Frequency (TF-IDF) can be used. Spatial information might be encoded so as to capture spatial reletions of the extracted features.

It has been shown that the vector quantization step is the computationally most expensive step in any Bag of Words framework. Fortunately, the majority of the cost is associated with the training part of the computation. When all the models are trained the only operation required in case of vector quantization is keypoint – visual vocabulary matching. Even so, diverse algorithms have been used in order to minimize the impact of clustering on the overall efficiency. The *approximate kMeans* is faster but insufficiently so as to solve the problem. Random Forests algorithm is considerably less expensive and can provide bettern Mean Average Precision (MAP) score.

Quite a different approach is using a Gaussian Mixture Model (GMM) algorithm. It is many times more expensive than kMeans, the tradeoff being higher precision. The GMM finds not only clusters' centroids but the gaussian distribiutions of points in each cluster. Therfore, in the keypoint-centroid matching stage a set of probabilities is obtained instead of a simple single-cluster assignment. These probabilities encode likelihoods of the keypoint belonging to each cluster. It is up to the user how to utilize this additional information.

4.4 Classification

Predicting a class associated with an image requires feeding the visual words histograms into a classifiers. A simple example of a classifier would be a K-nearest neighbours algorithm. In this setting every histogram is treated as a point in a multi-dimensional space. A class of an image is determined by classes of the nearest neighbours. Manifold of distinct metrics can be used: L1, L2 and others. The number of nearest neighbours (K) has to be determined.

More advanced classifiers can be divided into the two main classes: generative models and discriminative models.

Construction of a discriminative model requires a supervised learning approach. It can be seen as a learning by example method. The general aim is to compute an approximate mapping between representations of examples (the train set). The purpouse of this mapping is correct (e.g. with some success rate) prediction of labels for inputs which label is unknown. The model computation stage is called *training* of a model. After the model has been trained previously unseen examples can be fed into the classifier. The clasifier calculates some similarity measure between the unlabled input example and all the modeled classes. The resulting label is a label of which class had the highest score in terms of the similarity measure. The KNN is an example of a discriminative model. The most widely used discriminative classificator is a Support Vector Machine.

The generative models are usually Baysian text-based models. Many of them heavily depends on the Probabilistic Graphical Models concept. They include *Latent Semantic Analysis* with *probabilistic Latent Semantic Analysis*, *Latent Dirichlet Allocation* and others. These models allow discovery of topic distrubution in documents. In the image domain an image can be though of as a document and an object category as a topic. Under this conditions an image (and its category) can be modeled as a mixture of topics. Learning consists of finding topics distributions for each specific category.

References

[1] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, page 22, 2004.

[2] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524–531. IEEE, 2005.

[3] S. Filipe and L. A. Alexandre. A comparative evaluation of 3d keypoint detectors.

[4] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[5] X. Li and A. Godil. Investigating the bag-of-words method for 3d shape retrieval. *EURASIP Journal on Advances in Signal Processing*, 2010:5, 2010.

[6] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[8] J. Ponce, D. Forsyth, E.-p. Willow, S. Antipolis-Méditerranée, R. d'activité RAweb, L. Inria, and I. Alumni. Computer vision: a modern approach. *Computer*, 16:11, 2011.

[9] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 3212–3217. IEEE, 2009.

[10] R. B. Rusu, Z. C. Marton, N. Blodow, and M. Beetz. Persistent point feature histograms for 3d point clouds. *Intelligent Autonomous Systems 10: Ias-10*, page 119, 2008.

[11] S. Salti, F. Tombari, and L. D. Stefano. A performance evaluation of 3d keypoint detectors. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2011 International Conference on*, pages 236–243. IEEE, 2011.

[12] S. M. Smith and J. M. Brady. Susan—a new approach to low level image processing. *International journal of computer vision*, 23(1):45–78, 1997.

[13] R. Toldo, U. Castellani, and A. Fusiello. A bag of words approach for 3d object categorization. In *Computer Vision/Computer Graphics CollaborationTechniques*, pages 116–127. Springer, 2009.

[14] C.-F. Tsai. Bag-of-words representation in image annotation: A review. *ISRN Artificial Intelligence*, 2012, 2012.

[15] K. E. van de Sande, T. Gevers, and C. G. Snoek. Empowering visual categorization with the gpu. *Multimedia, IEEE Transactions on*, 13(1):60–70, 2011.

[16] Q. Zhang, X. Song, X. Shao, R. Shibasaki, and H. Zhao. Category modeling from just a single labeling: Use depth information to guide the learning of 2d models. In *Computer Vision and Pattern Recognition, 2013. CVPR 2013. IEEE Computer Society Conference on*. IEEE, 2013.

[17] Y. Zhong. Intrinsic shape signatures: A shape descriptor for 3d object recognition. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 689–696. IEEE, 2009.