Research Proposal: Fast and Accurate Recognizing Human Actions in Massive Still Images

Wenrui Diao * Qian Zhang † Haichao Zhu ‡

October 3, 2013

1 Introduction

Recently, the computer vision field has made great progress in recognizing isolated objects, such as faces and cars. But a large proportion of our visual experience involves recognizing the human action. Human action is an essential part of many images. Recognizing human actions in still images will lay the foundations to many applications such as image retrieval from large-scale image archives. Besides that, human action recognition can also help solving other problems for still images such as object recognition or scene estimation. Recognition of human actions has mostly been explored in video. While the motion of people often provides discriminative cues for action classification, actually many actions can be identified from single images, since humans can recognize action based on only static images.

The goal of our work is to use some **big data techniques** to recognize scenes in which a person is interacting with a specific object in a specific manner, such as playing musical instruments in still images. Our intuitive approach for this problem is to use the traditional visual data mining. The standard flowchat of visual data mining is first to do feature extraction, and then clustering [1] on this features to generate code books. For object recognition tasks, nearest neighboring search or some machine learning methods will be used. However the detection rates are still very low. Inspired by the idea from [2], we will also design an Locality-sensitive hashing [1] scheme to accelerate the K-nearest neighbor search in object detection.

2 Literature Review

There are a wide range of approaches to human action recognition in the community of computer vision. Analyzing human dynamics from video is a common theme to many of these approaches. However, recognizing human

*SID: 1010089510, Email: dw013@ie.cuhk.edu.hk
†SID: 1155005691, Email: qzhang@cse.cuhk.edu.hk
‡SID: 1010087980, Email: hczhu@cse.cuhk.edu.hk

actions in still images has received little attention with the exception of few papers [3–7]. The methods proposed in [3, 5, 7] mainly relied on the cues of body pose for recognition.

To avoid reasoning about human body poses, the bag-of-features (BoF) and the spatial pyramid matching (SPM) are used in [8]. The BoF [9] based SPM has two important steps. First, features points are detected on the input images, and then descriptors are extracted from each feature point. Next, a code book is applied to quantize each descriptor and generate the "code". Generally, the code book and the codes of descriptors are generated at the same time, like K-means.

There are many methods to extract image features in the computer vision fields, such like SIFT [10], HOG [11]. And those robust low-level image features have been proven to be effective representations for a variety of visual recognition tasks such as object recognition and scene classification. Besides these low level features, some high level image features are proposed, like Object Bank [12], Action Bank [13]. These high level features are usually build on the top of the low level features. They are also robust and have more semantic meaning. A lot of researches have been done to apply these features into the SPM framework.

To improve the scalability of SPM, several methods are proposed to obtain the codes. In particular, Yang et al. [14], proposed the ScSPM method where sparse coding (SC) was used instead of K-means to obtain nonlinear code. Wang et al. [15] proposed a modification to SC, called Local Coordinate Coding (Locality-constrained Linear Coding). Both of these two methods achieves state-of-art performances on several benchmarks.

In the aspect of searching similarity items, several cluster algorithms consider information of all dimensions at the same time and try to utilize information as much as possible. But with the data points expansion in a high dimensional space, the distance difference will become small. So similarity and dissimilarity between two points will become meaningless [16]. Some methods have been proposed for overcoming these challenges using dimension reduction or subspace comparison. LSH could be treated as a good solution performing probabilistic dimension reduction of high-dimensional data [17].

3 Methodology

The traditional method [9] to solve such scene classification is first to do feature extraction and then apply clustering on these features. For feature extraction part, we will utilize Object Bank method [12] to extract high level features. Then we will build code book on the top of features using Locality-constrained Linear Coding [14]. To speed up the K-nearest neighbor search, we will also implement a locality-sensitive hashing schema.

3.1 Object Bank Method

Low-level image features have been proven to be effective representations for a variety of visual recognition tasks such as object recognition and scene

classification; but pixels, or even local image patches, lack of semantic meanings. However, for high level visual tasks, such low-level image features are potentially not enough. Thus in this work, we consider to use a high-level image representation called Object Bank, first proposed in [12]. In Object Bank, an image is represented as a scale-invariant response map of a large number of pre-trained generic object detectors, blind to the testing dataset or visual task.

The traditionally methods in SPM [9, 14, 15] are usually dominated by the low level features, such like SIFT [10], HOG [11]. However, in our work, we will use the Object Bank method to extract the high level features instead of low level ones. That is mainly because **human actions have semantic meanings** [7]. The same poses can have different meanings based on the context. So we need high level semantic information. For example, given two images with a person running. However, in the first image it also has a football in the front of the person, while on the other image, there is not. So the first image will be interpreted as kicking while the second one will be regarded as running.

3.2 Locality-constrained Linear Coding

Locality-constrained Linear Coding (LLC) [15] is an effective coding scheme, which is designed to take the place of K-means in traditional Spatial Pyramid Matching (SPM). LLC utilizes the locality constraints to project each descriptor into its local-coordinate system, and the projected coordinates are integrated by max pooling to generate the final representation. LLC performs significantly better than the traditional nonlinear SPM, due to the linear classifier. Besides that, LLC also has an analytical solution compared with sparse coding schema [14]. In addition, LLC also has a fast approximated method using K-nearest neighbor searching.

Generally, the LLC coding is used to generate the code book on low level features. In our work, We will build the code book on the top of high level features using Object Bank. To the best of our knowledge, there is no paper about using LLC on Object Bank features.

3.3 Locality-sensitive Hashing

Locality-sensitive hashing (LSH) algorithm [17,18] is a kind of technique that allows one to quickly find high dimensional similar entries in large databases. This algorithm does not guarantee an exact solution, however it will either return the correct answer or a close one with high probability.

The key idea is to hash the points using several hash functions, to ensure that for each function, the probability of objects are close to each other is much higher than the probability of objects are far apart. Then, one can determine near neighbors by hashing the query point, and retrieving elements stored in buckets containing that points. The features of LSH meet our needs for similarity search.

3.4 Datasets

We will consider two datasets in our work.

First is the "People Playing Musical Instrument (PPMI)" [7], a dataset of human and object interaction activities in our projects. It has images of human interacting with twelve different musical instruments. They are: bassoon, cello, clarinet, erhu, flute, French horn, guitar, harp, recorder, saxophone, trumpet, and violin. Each class includes 200 PPMI+ images (humans playing instruments) and 200 PPMI- images (humans holding the instruments without playing). Totally it has 4800 images in 800 megabytes.

The second dataset is from [3], focusing on sports. It contains six classes. They are cricket batting, cricket bowling, croquet, play with tennis forehand, tennis with serve and smashing volleyball. Each class contains 50 images.

Our work mainly focus on two problems. First is to recognize different human action among this two datasets. Besides, we will also try to distinguish PPMI+ images with PPMI- images. The second one is much more challenge than the first one.

4 Timetable

Our schedule could be divided into 3 stages. The first stage (half month) will focus on demand analysis and scheme design. We will finish code implementations and case-tests in the following stage, which will cost about one month. The last stage (middle November) contains the tasks of report writing and presentation preparation. Also this timetable may be adjusted according to the actual processes.

5 Conclusion

This research proposal describes a scheme for fast and accurate recognizing human actions in massive still images. The to-do theory analysis and testing platform building are also provided. The expected successful result is that we could execute real test cases with high performance and accuracy.

References

[1] A. Rajaraman and J. D. Ullman, *Mining of massive datasets*. Cambridge University Press, 2012.

[2] T. Dean, M. A. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, and J. Yagnik, "Fast, accurate detection of 100,000 object classes on a single machine," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[3] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 10, pp. 1775–1789, 2009.

[4] N. Ikizler, R. G. Cinbis, S. Pehlivan, and P. Duygulu, "Recognizing actions from still images," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.

[5] Y. Wang, H. Jiang, M. S. Drew, Z.-N. Li, and G. Mori, "Unsupervised discovery of action classes," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 1654–1661.

[6] L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[7] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 9–16.

[8] V. Delaitre, I. Laptev, and J. Sivic, "Recognizing human actions in still images: a study of bag-of-features and part-based representations." in *BMVC*, vol. 2, no. 5, 2010, p. 7.

[9] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 2169–2178.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[12] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Advances in neural information processing systems*, 2010, pp. 1378–1386.

[13] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1234–1241.

[14] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1794–1801.

[15] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3360–3367.

[16] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: a review," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 90–105, 2004.

[17] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the thirtieth annual ACM symposium on Theory of computing*. ACM, 1998, pp. 604–613.

[18] A. Andoni and P. Indyk, "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions," in *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*. IEEE, 2006, pp. 459–468.