Superpixel Group Mining for Manipulation Action Recognition

Tianjun Huang and Stephen McKenna

CVIP, School of Science and Engineering,
University of Dundee, Dundee DD1 4HN, UK
{t.huang, s.j.z.mckenna}@dundee.ac.uk

Abstract. Manipulation action recognition is a challenging problem in computer vision. We previously reported a system based on matching groups of superpixels. In this paper, we modify the superpixel group mining algorithm and report results on two datasets. Recognition accuracies are comparable with those reported using deep learning. The representation used in our approach is amenable to interpretation. Specifically, visualisation of matched groups provides a level of explanation for recognition decisions and insights into the likely generalisation ability of action representations.

Keywords: Superpixel group mining · Action recognition · Computer vision.

1 Introduction

Manipulation actions usually contain fine-grained motions involving both actor and manipulated objects, in contrast to actions such as, e.g., running and jumping. One approach to recognition of manipulation actions is to build object and human body part detectors and analyse the relationships between them [6, 1, 2]. However, supervised training of detectors for all objects of interest can require extensive manual image annotation [1–5]. Another shortcoming is that object transformations arising from manipulation are not always sufficiently represented [4, 14]. Actions such as those involved in food preparation can markedly change object appearance (e.g., mixing ingredients) and topology (e.g., cutting into pieces). This situation is not well-handled by spatial-temporal tube methods for example. Some other methods relied on pose trackers (e.g., [6]) and assumed that most of the human body appears in the camera view. Yang et al. [7] used an unsupervised method to segment objects for recognising manipulation actions against a clear, uncluttered background.

We proposed an action recognition system based on discriminative superpixel group mining which avoids the need for manual object annotations and which can represent object transformations [8]. However, in order to select the best representations for each action, the representative property should be considered as well in the mining process [9]. In this work, we modify the discriminative group mining algorithm by including representativity. We report results on two

datasets: 50 Salads [10] and Actions for Cooking Eggs (ACE) [11]. We illustrate that the method learns representations that are amenable to interpretation via visualisation, providing insights into recognition decisions and generalisation.

2 Proposed Method

2.1 On-line Spatio-temporal Superpixel Grouping

We briefly introduce our superpixel grouping algorithm. More details can be found in [8]. Each frame is first over-segmented into superpixels by using Depth SEEDS [12]. RANSAC [13] is applied to find the plane of work surface. Superpixels above this surface are connected spatially and temporally based on color similarity and optical flow to sequentially build spatio-temporal superpixel groups. These groups can contain temporal bifurcations and loops so that they are able to represent complex object transformations in actions such cutting and mixing.

2.2 Group Representation and Matching

We use colour, motion and texture to represent each superpixel group. Colour is represented by a histogram (25 bins per channel), motion by an optical flow orientation histogram weighted by flow magnitudes (30 bins), and texture by a histogram of oriented gradients (30 bins). Let $a(g_i, g_j)$, $m(g_i, g_j)$ and $h(g_i, g_j)$ denote repectively the intersections of the colour, flow, and texture histograms of two superpixel groups g_i and g_j . These groups' similarity $k(g_i, g_j)$ is computed as in Eqn. (1) where $\beta_3 = 1 - \beta_1 - \beta_2$ and the parameters β_1 and β_2 are tuned during the training process.

$$k(g_i, g_j) = \beta_1 a(g_i, g_j) + \beta_2 m(g_i, g_j) + \beta_3 h(g_i, g_j) \tag{1}$$

2.3 Mining and Recognition

Previously [8], mining used the seeding algorithm in [14] which considers discriminability. Here we also include representativeness in the mining process [9]. The idea is that, for example, mined superpixel groups for *action A* should only tend to appear in instances of *action A* (disciminability) and that they should appear in many instances of *action A* (representativeness). To achieve this, for each group g_i in the training set, we select the M most similar groups from each different subject who performed the manipulation action. The total number of selected groups is then $K = M \times (P - 1)$ where P is the number of subjects in the training set. We compute the mining score for a group by summing its discriminability and representativeness scores. The former is the proportion of selected groups with the same label as that group. The latter is the proportion of subjects with at least one selected group with the same label.

A video frame is assigned an action label based on a fixed duration temporal window centred on that frame. Max-N pooling [14] is used to generate the feature vector for a temporal window. Implementation details can be found in [14, 8]. Windows are classified using support vector machines trained in LibLinear [15].

3 Experiments

We used two datasets: 50 Salads and ACE. The 50 Salads dataset contains 50 videos. It has 25 subjects; each subject made two mixed salads with non-unique order of steps. There are 10 actions in this dataset: *add pepper, add oil, mix dressing, peel cucumber, cut ingredient, place ingredient into bowl, mix ingredients, serve salad onto plate, dress salad and NULL*, where *NULL* represents all times when one of those 9 actions is not occurring. Following the protocol in [10], the dataset is split into 5 folds. Each fold contains 10 videos made by 5 subjects. Five-fold cross validation is used to estimate performance.

The ACE dataset was proposed in the contest "Kitchen Scene Context based Gesture Recognition" at ICPR 2012. There are seven subjects. Each of them was required to cook five recipes. Nine actions were annotated in the dataset: *breaking, mixing, baking, turning, cutting, boiling, seasoning, peeling and NULL*, where *NULL* represents all times when one of those 8 actions is not occurring. There are 25 videos in the training set and 10 videos in the testing set.

We randomly selected 10,000 superpixel groups from 4,000 temporal windows in each action class for group mining. Each temporal window has a duration of 155 frames.

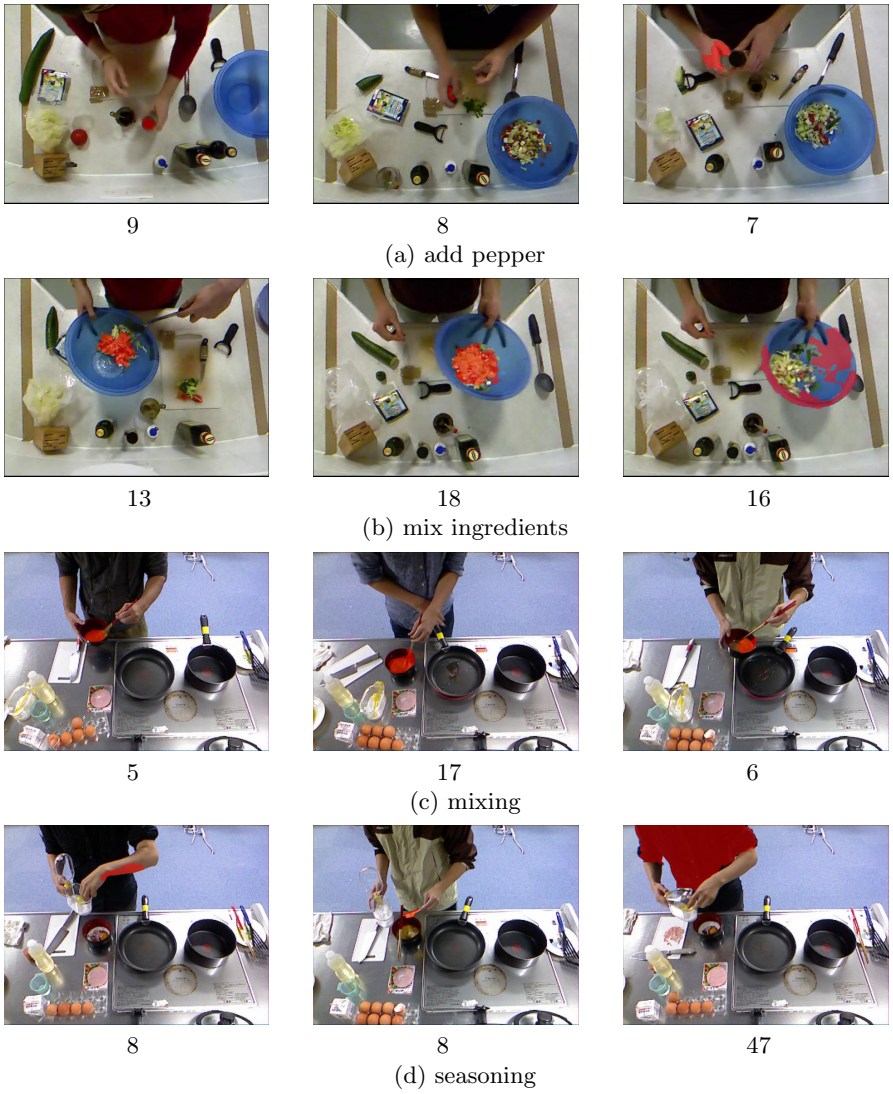
Fig. 1 shows examples of mined superpixel groups; red regions are superpixels in the mined groups. The mined groups provide interpretable representations for the different actions. For instance, in the 50 Salads examples, groups representing the pepper container and the hand motion suggest the action *add pepper*; groups representing food ingredients with groups on the bowl suggest the action *mix ingredients*. In ACE examples, mined groups capture the eggs in the bowl as the representation for action *mixing*; superpixel groups of human arm and spoon together suggest the action *seasoning*.

By visualising mined groups, we can discover if they provide a representation that is likely to generalise. For instance, the third group in Fig. 1(d) *seasoning* captures clothing rather than anything inherently associated with the action class. This indicates overfitting and may cause failure to generalise.

Table 1 compares the modified mining method with the previous method [8] on both datasets. Accuracies on 50 Salads are similar. As reported in [8] this accuracy is better than that of competing methods using deep learning. The modified mining method improved accuracy a little on the ACE dataset.

Table 1. Measurements on two datasets.

50 Salads	Precision	Recall	F_1 score	Frame-wise accuracy
Superpixel Group Mining [8]	66	68	67	76.5
Modified Group Mining	**66**	**69**	**67**	**76.6**
ACE	Precision	Recall	F_1 score	Frame-wise accuracy
Superpixel Group Mining [8]	65	66	64	68.2
Modified Group Mining	**70**	**71**	**68**	**71.6**



9 8 7
(a) add pepper

13 18 16
(b) mix ingredients

5 17 6
(c) mixing

8 8 47
(d) seasoning

Fig. 1. Examples of mined superpixel groups (red regions). Numbers are durations of groups in frames. (a-b) 50 Salads dataset. (c-d) ACE dataset.

4 Conclusion

We modified the superpixel group mining used in our previously proposed method for manipulation action recognition. Experiments on two datasets showed the effectiveness in terms of accuracy. We also highlighted the interpretable nature of the learned representation, in contrast to many deep learning methods for example. Visualisation of matched superpixel groups can provide a level of explanation

for the recognition decisions made. It can also provide insights into likely generalisation ability, enabling identification of groups that represent aspects of the video that are not relevant to the actions of interest.

References

1. A. Prest, V. Ferrari, C. Schmid: Explicit modeling of human-object interactions in realistic videos. IEEE Trans. PAMI, 835–848 (2013)
2. Y. Zhou, B. Ni, S. Yan, P. Moulin, Q. Tian: Pipelining localized semantic features for fine-grained action recognition. In ECCV (2014)
3. B. Ni, V. R. Paramathayalan, T. Li, P. Moulin: Multiple granularity modeling: a coarse-to-fine framework for fine-grained action analysis. Int. J. Computer Vision, 28–43 (2016)
4. B. Ni, X. Yang, S. Gao: Progressively parsing interactional objects for fine grained action detection. In CVPR (2016)
5. C. Fermüller, F. Wang, Y. Yang, K. Zampogiannis, Y. Zhang, F. Barranco, M. Pfeiffer: Prediction of Manipulation Actions. Int. J. Computer Vision, 1–17 (2017)
6. B. Packer, K. Saenko, D. Koller: A combined pose, object, and feature model for action understanding. In CVPR (2012)
7. Y. Yang, C. Fermuller, Y. Aloimonos: Detection of manipulation action consequences (MAC). In CVPR (2013)
8. T. Huang, S. J. McKenna: Sequential recognition of manipulation actions using discriminative superpixel group mining. In ICIP (2018)
9. B. Fernando, E. Fromont, T. Tuytelaars: Mining mid-level features for image classification. Int. J. Computer Vision, 186–203 (2014)
10. S. Stein, S. J. McKenna: Combining embedded accelerometers with computer vision for recognizing food preparation activities. In ACM UbiComp (2013)
11. A. Shimada, K. Kondo, D. Deguchi, G. Morin, H. Sterna: Kitchen scene context based gesture recognition: a contest in ICPR2012. Advances in Depth Image Analysis and Applications, 168–185 (2013)
12. M. Van den Bergh, D. Carton, L. Van Gool: Depth SEEDS: Recovering incomplete depth data using superpixels. In WACV (2013)
13. M. A. Fischler, R. C. Bolles: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Comm. ACM, 381–395 (1981)
14. Y. Zhou, B. Ni, R. Hong, M. Wang, Q. Tian: Interaction part mining: a mid-level approach for fine-grained action recognition. In CVPR (2015)
15. R. Fang, K. Chang, C. Hsieh, X. Wang, C. Lin: LIBLINEAR: a library for large linear classification. J. Machine Learning Research, 1871–1874 (2008)