Recognizing Human Activity in Still Images by Integrating Group-Based Contextual Cues

Zheng Zhou
Beijing Institute of Technology
zz24@bit.edu.cn

Kan Li
Beijing Institute of Technology
likan@bit.edu.cn

Xiangjian He
University of Technology, Sydney
xiangjian.he@uts.edu.au

ABSTRACT

Images with wider angles usually capture more persons in wider scenes, and recognizing individuals' activities in these images based on existing contextual cues usually meet difficulties. We instead construct a novel group-based cue to utilize the context carried by suitable surrounding persons. We propose a global-local cue integration model (GLCIM) to find a suitable group of local cues extracted from individuals and form a corresponding global cue. A fusion restricted Boltzmann machine, a focal subspace measurement and a cue integration algorithm based on entropy are proposed to enable the GLCIM to integrate most of the relevant local cues and least of the irrelevant ones into the group. Our experiments demonstrate how integrating group-based cues improves the activity recognition accuracies in detail and show that all of the key parts of GLCIM make positive contributions to the increases of the accuracies.

Categories and Subject Descriptors

U.3 [**Understanding**]: Multimedia and Vision

Keywords

activity recognition, context, group-based cue, Fusion-RBM, focal subspace measurement

1. INTRODUCTION

Image-based human activity recognition has attracted increasing research attention in very recent years. It is of great scientific importance and has useful applications in multimedia, such as image annotation, behavior based image retrieval, video frame reduction and human computer interaction. An activity contains a number of subsequent actions and gives an interpretation of the movement that is being performed [8]. Unlike video-based methods that model the subsequent actions by spatio-temporal features, image-based methods utilize contextual cues to help characterize an activity because there is no temporal information available in still images.

The existing popular contextual cues include action-related object cues, human-object interaction cues, and whole scene cues [5].

MM '15, October 26-30, 2015, Brisbane, Australia.
 ISBN 978-1-4503-3459-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2733373.2806300.

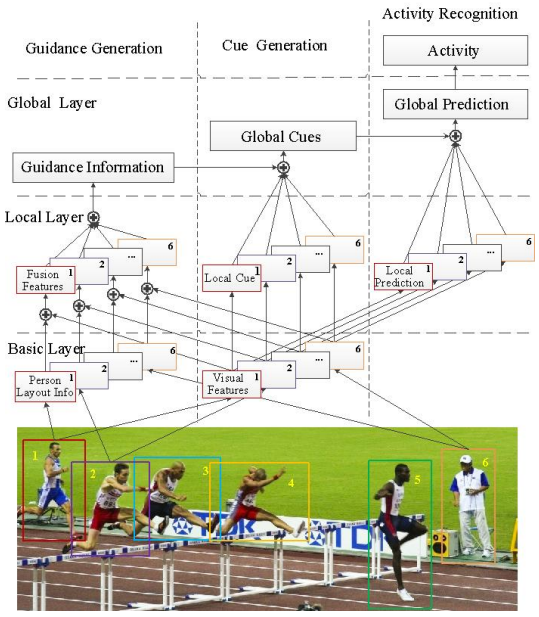


Figure 1: GLCIM for Human Activity Recognition.

Action-related object cues and human-object interaction cues work well when the "key object" can easily be found and recognized [9]. A scene cue models the occurrence of a scene, an object and an activity, and it is very suitable for the scene-specific activities [7]. However, for images with wider angles, taken in daily life and usually containing more persons, applying the above-mentioned three kinds of cues may lead to the difficulties as follows. "Key objects" are illegible and hard to be detected and recognized in a complex background, and the same scene contributes very little to discriminate persons with different activities in an image.

As a matter of fact, when humans see an image containing multiple persons, they confirm the activity of a certain person by referring to a group of persons who perform the same activity or related activities. In this paper, we try to model this biomimetic mechanism in computer vision and we call it a group-based contextual cue. For example, in Fig.1, person 1 is hurdling, but the action or pose of him looks more like running. If we apply a group-based cue by referring to persons 2-5, we will be sure of telling that person 1 is hurdling. An action-related object cue and a human-object interaction cue will work worse in this image because person 1 has no interaction with these hurdles. Scene cue will assertively tell that person 6 is hurdling too, but he is only a referee. This problem can also be solved in the following two procedures in theory: discovering groups and recognizing group activities. However, persons performing the same group activity may do different individual

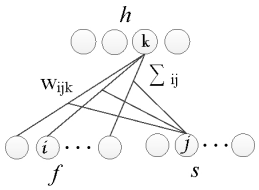


Figure 2: The Structure of FRBM.

activities. For example, two persons as a group working collaboratively to take photos may perform different individual activities, e.g., operating a camera and posing.

Two main challenges exist in generating group-based cues. One is taking full advantage of the relevant contextual information, and the other is minimizing the effect of the irrelevant or "misleading" contextual cues. In our global-local cue integration model, we design a series of methods and algorithms to select cues carried by suitable persons, including a fusion restricted Boltzmann machine (FRBM) to fuse features and a focal subspace measurement together with a global cue integration method based on entropy to integrate the most relevant local cues and stop the integrating process when misleading cues try to decrease the significance of the integrated global cues.

We highlight the main contributions of this paper as follows.

- We use a group-based cue to help human activity recognition in still images. It is for the first time that context carried by surrounding persons is formed as a contextual cue to do the activity recognition for a single person in an image.
- We develop a global-local cue integration model (GLCIM) for recognizing a human activity using a group-based cue. GLCIM can select suitable local cues to generate reasonable global cues which integrate least of the irrelevant context.
- We also propose a fusion restricted Boltzmann machine (FRBM) and a corresponding focal subspace measurement to estimate the interdependencies of persons. They play important roles in GLCIM and also allow the model to work well when the amount of training data is limited.

2. RELATED WORK

The existing and the most popular contextual cues for human activity recognition based on images are action-related object cues [9], human-object interaction cues [12], and whole scene cues [11]. An action-related object cue helps activity recognition by discovering objects that may relate to the activity. A human-object interaction cue further considers the relative position or relative angle between an object and a human to disambiguate some object-inferred activities. A scene cue utilizes the fact that some activities are happening in specific places. Some existing work jointly employs these cues. For example, Li et al. [7] combined object cues and scene cues to model sports events and Shapavolavo et al. [10] integrated all three kinds of cues to beat the state-of-the-art results on the PASCAL dataset. Unlike their approaches, which extract context from objects and scenes, we exploit the contextual information provided by surrounding persons instead, and our model can be easily extended to these three kinds of cues.

There is also a line of work on modeling group activities. Quite a proportion of them focus on video data [2]. Some of them also employ complex equipment, such as multiple cameras [13]. For the existing research on group activity recognition based on still images [1] [6], researchers pay much attention to how to model a group activity based on the activities of individual members. Moreover, all persons in one image share the same activity label in the group activity recognition. However, our work is to recognize the activities of individuals and the persons in one image can do different activities. A group-based cue is to find a group of persons that may provide useful and contextual information to the interested person. Persons doing the same activity may have other groups of persons that provide the cues.

3. GLOBAL-LOCAL CUE INTEGRATION MODEL

Global-local cue integration model (GLCIM) is proposed to select context carried by suitable persons and form a global cue based on the selected local cues to help the activity recognition.

3.1 Overview

The overview of GLCIM is shown in Fig.1. A local cue extracted from a surrounding person is a vector of probabilities that describe how much evidence that this surrounding person provides to infer an interested person doing potential activities. A global cue is generated by a combination of a group of selected local cues. Guidance information reflects the interdependencies of pairs of persons and determines which persons' local cues are selected. We assume the interdependencies are related to two key points: 1) the layout information of persons, and 2) the similarities in persons' visual features, such as clothes colors and poses. We design a fusion-restricted Boltzmann machine (FRBM) to form the fusion features based on the visual features extracted by a convolutional neural network (CNN) and the layout information. A focal subspace measurement is also developed to measure the relative distances between the fusion features of pairs of persons, and this allows the GLCIM to work with a limited amount of training data. With the guidance information, a proper group of local cues are chosen to form the global cue. Once the global cue for an interested person is generated, we can use it to improve the local activity recognition.

3.2 Fusion Restricted Boltzmann Machine

To fuse the visual features and the layout information, which are quite different in scale, we design a FRBM. The structure of the FRBM is shown in Fig.2. f and s are two visible layers standing for the visual features and the layout information respectively, and h is the hidden layer, in which the values are deemed as fusion features. FRBM is trained by minimizing the normalized reconstructing error of visible layers, so the generated fusion features are not influenced by the different scales of original features.

The energy function of FRBM capturing all possible correlations among the components of the visible f, s and the hidden h is given as:

$$E(f,s,h) = -\sum_{ijk} w_{ijk} f_i s_j h_k - \sum_i w_i^f f_i - \sum_j w_j^s s_j - \sum_k w_k^h h_k, \quad (1)$$

where $w_{ijk} \in W$ is the connection weight, $w_i^f \in W^f$ is the bias of layer f, $w_j^s \in W^s$ is the bias of layer s, and $w_k^h \in W^h$ is the bias of layer h. We define the parameter set $\theta := (W, W^f, W^s, W^h)$. Then, we have the joint probability distribution:$p(f,s,h) = \frac{e^{-E(f,s,h)}}{\sum_{f,s,h} e^{-E(f,s,h)}}$.

By maximizing the likelihood, we get the weight update rule: $\Delta\theta = \varepsilon \cdot \frac{\partial \ln p(f,s)}{\partial \theta}$,where ε is the learning rate. Then, for updating each parameter, we calculate the partial derivative of each component of θ. With these partial derivatives, we calculate the hidden values of h by Gibbs sampling as follows. First, the hidden units are sampled according to $p(h_k = 1|f,s) = sigmoid(\sum_{ij} w_{ijk} f_i s_j + w_k^h)$, and then the units of f and s are sampled through iteration between

$$\begin{cases} p(f_i = 1|s,h) = sigmoid(\sum_{jk} w_{ijk} s_j h_k + w_i^f) \\ p(s_j = 1|f,h) = sigmoid(\sum_{ik} w_{ijk} f_i h_k + w_j^s). \end{cases}$$

Figure 3: Examples in our dataset.

3.3 Focal Subspace Measurement

A FRBM can be regarded as a mapping that maps data points from original feature spaces to a unified fusion feature space. We need plenty of data to train the mapping between the spaces with large sizes. However, in GLCIM, we only need the relative distances from the surrounding persons to one certain interested person. Therefore, we propose the focal subspace measurement to give the relative distances. The focal subspace distance is defined as:

$$FSDis_{p_f}(p_i,p_j) = EuDis(Tran_{p_f}(p_i), Tran_{p_f}(p_j)), \quad (2)$$

where p_f is the original data of the interested focal person, $Tran_{p_f}(p_i)$ means the hidden layer values when the data of the person p_i is run through the FRBM trained only by the data of p_f, and $EuDis(\cdot)$ stands for the Euclidean distance.

3.4 Global Cue Generation

A global cue is in the form of a vector of probabilities like a local cue. The guidance information is about the distances from the surrounding persons to the interested person measured by the focal subspace measurement, which reflects the possibilities of persons' doing the same or related activities. The higher possibility a surrounding person is with, the higher contribution he will make in deducing the activity of the interested person. Thus, we construct the global cue by accumulating the local cues one by one in the order of the possibilities from highest to lowest with the inverse of the entropies of local cues as weights, and we stop the accumulation when the significance of the accumulation result reaches the peak. The significance or stability of a set of possibilities is measured by its entropy, defined as: $Ent(P) = -\sum_P p \ln p$. In this way, the most relevant cues measured by the focal subspace measurement will be integrated and when irrelevant cues are going to cause the fall of the $Ent(P)$, the integrating process will stop.

Finally, the global prediction is given by:

$$GP(p_f) = C \cdot \left(\frac{LP(p_f)}{Ent(LP(p_f))} + \frac{GC(p_f)}{Ent(GC(p_f))}\right), \quad (3)$$

where $C = \left(\frac{1}{Ent(LP(p_f))} + \frac{1}{Ent(GC(p_f))}\right)^{-1}$ is a normalization constant, and $LP(p_f)$ and $GC(p_f)$ are vectors of possibilities standing for the local prediction and the global cue of the focal person p_f respectively. $LP(p_f)$ can be obtained by any existing activity recognition method as long as the method can give the possibilities of the person performing the potential activities.

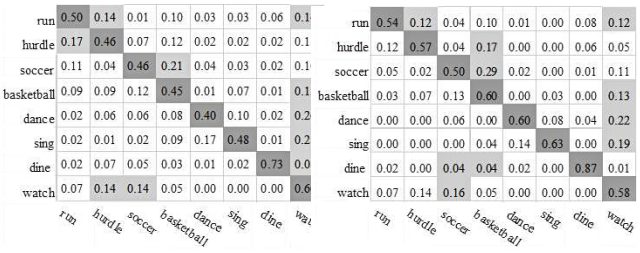
4. EXPERIMENTS

Many public datasets are available to validate human activity recognition methods based on still images. However, most of them are collected for classifying the activity of a single person. In datasets for pair of persons' interactions and collective activities of larger groups, only one group activity is label for each image. Each image in Structured Group Dataset (SGD) [3], a very recently proposed and challenging dataset, contains groups of persons, but only a small proportion of the images contain groups performing different activities. This dataset is not sufficient for our testing.

In this paper, we compile a new dataset with two characteristics described as follows. 1) Each image in the dataset contains several persons. 2) Persons in an image may execute different activities. Persons in this dataset are doing 8 activities: running, hurdling, playing soccer, playing basketball, dancing, singing, dining and watching. Watching is a relatively loose activity, and it can be an audience staring at a running race or a referee supervising a basketball game. We assign the watching label to the persons whose activities are different from the main activity in an image. In each activity category, there are about 2000 persons being detected. As shown in Fig.3, this dataset is a very challenging one and we highlight the difficulties as follows. 1) The activity classes are diverse, and some classes, such as singing and dancing, are hard to discriminate. 2) Within the same activity, the sizes and poses of person instances are very different. 3) The background of each image is highly cluttered and diverse. 4) Except the activity labels, no any other information, including basic segmentation, is given.

Table 1: Comparison of Activity Recognition Accuracies.

	Without Cue	With Cue
total	50.83%	61.15%
running	49.56%	53.98%
hurdling	46.43%	57.14%
soccer	46.49%	50.00%
basketball	44.78%	59.70%
dancing	40.00%	60.00%
singing	48.28%	62.93%
dining	72.55%	87.25%
watching	60.47%	58.14%

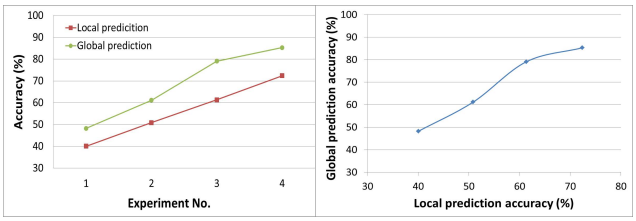


(a) (b)

Figure 4: The confusion matrices of the baseline (a) and our method (b).

To prove that our GLCIM can improve the activity recognition performance by integrating a group-based cue and to validate the performance of select suitable local cue, we first need local predictions and local cues. Since no detailed segmentation is contained in our dataset, heavy pre-process is needed if we employ existing methods on recognizing the activity of a single person. Therefore, we choose feed-forward neural networks to give both the local cues and the local predictions. For the local cues, we train a neural network by the images in our dataset and the statistical data of activity labels. For the local prediction, tuning the amount of total number of the nodes in the neural network allows us to simulate local predictions with different accuracies. We employ the part-based deformable model [4] to detect the persons in an image and the C-NN to extract features from the detected persons. We implement the whole process from the raw images to the final prediction result automatically on this very challenging dataset. Our approach is a pioneering work working on the activity recognition of a person using the cues of other persons' activities in the same group and is very different from the existing methods, so it cannot be compared with the existing methods directly.

Table 1 shows the comparison of our activity recognition method with a group-based cue and the baseline without any cues at the 50% local prediction accuracy. We can see that except for the



(a) (b)

Figure 5: The activity recognition accuracies of our activity recognition based on different local prediction accuracies.

watching activity, the accuracies of recognizing various activities increase between 3.51% to 20%. The confusion matrices in Fig.4 indicate the soundness of the increases, which benefit from 1) discriminating closer activities (referring to running and hurdling) and 2) removing unrelated activities (referring the increase of singing). For the case of watching activity, we see the decrease of accuracy as shown in Table 1. It is because, in some images as shown in Fig.1, the persons labeled watching are so isolated that all cues in the images have negative effects in recognizing the watching activities of the persons. GLCIM tries to select no local cues in this situation. Therefore, we have been happy enough to see this very small 2.33% decrease and it proves that GLCIM has integrated only very minimum number of mistaken local cues.

Fig.5 shows the activity recognition accuracies of GLCIM based on different local prediction accuracies. We can see that, for different local prediction results, we always achieve a significant improvement, and the increase of the global prediction accuracy shows an upward tendency.

Fig.6 shows a series of experiments to test the key parts in GLCIM. From (a), we see that the fusion features are a valid mix of the original features. From (b), we see that, with no extra training data, the focal subspace measurement outperforms the method of training a FRBM mapping for measurement. From (c), we see that, when many persons appear and interfere each other in one image, GLCIM achieves better performance than clustering the persons in their original feature spaces and combining all persons for a cue.

5. CONCLUSION

In this paper, we have presented a novel global-local cue to help recognize human's activities in still images. We have proposed a global-local cue integration model (GLCIM) to form the group-based cue based on context carried by suitable persons. We have also proposed a fusion restricted Boltzmann machine and a focal subspace measurement to estimate the interdependency between pairs of persons even if the amount of training data is limited. Our experimental results have demonstrated that the accuracies of recognizing human activities in still images are improved by integrating group-based cues and all key parts of GLICM contribute positively to the accuracy increases.

6. ACKNOWLEDGEMENTS

This work was supported by National High Technology Research and Development Program of China (No.2013CB329605) (973 Program) and Training Program of the Major Project of BIT.

References

[1] AMER, M. R., XIE, D., ZHAO, M., TODOROVIC, S., AND ZHU, S.-C. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In *ECCV, 187-200, 2012*.

[2] CHENG, Z., QIN, L., HUANG, Q., YAN, S., AND TIAN, Q. Recognizing human group action by layered model with multiple cues. *Neurocomputing 136* (2014), 124–135.

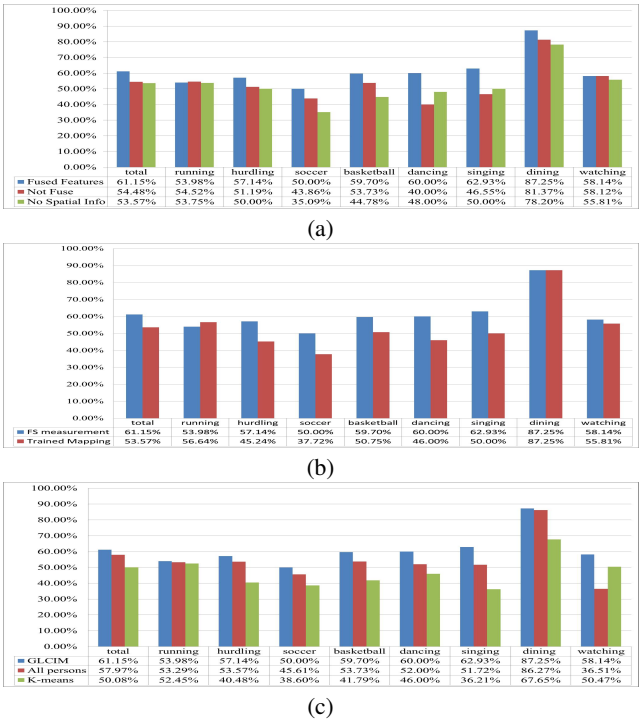


Figure 6: Experiments on the effect of the details in global cue generation. (a) Validation for FRBM. (b) Validation for focal subspace measurement. (c) Validation for local cue selecting algorithm.

[3] CHOI, W., CHAO, Y.-W., PANTOFARU, C., AND SAVARESE, S. Discovering groups of people in images. In *ECCV, 417-433, 2014*.

[4] FELZENSZWALB, P. F., GIRSHICK, R. B., MCALLESTER, D., AND RAMANAN, D. Object detection with discriminatively trained part-based models. *T-PAMI 32*, 9 (2010), 1627–1645.

[5] GUO, G., AND LAI, A. A survey on still image based human action recognition. *Pattern Recognition* (2014).

[6] LAN, T., WANG, Y., YANG, W., ROBINOVITCH, S. N., AND MORI, G. Discriminative latent models for recognizing contextual group activities. *T-PAMI 34*, 8 (2012), 1549–1562.

[7] LI, L.-J., AND FEI-FEI, L. What, where and who? classifying events by scene and object recognition. In *ICCV, 1-8, 2007*.

[8] MOESLUND, T. B., HILTON, A., AND KRÜGER, V. A survey of advances in vision-based human motion capture and analysis. *CVIU 104*, 2 (2006), 90–126.

[9] PREST, A., SCHMID, C., AND FERRARI, V. Weakly supervised learning of interactions between humans and objects. *T-PAMI 34*, 3 (2012), 601–614.

[10] SHAPOVALOVA, N., GONG, W., PEDERSOLI, M., ROCA, F. X., AND GONZÀLEZ, J. On importance of interactions and context in human action recognition. In *Pattern Recognition and Image Analysis*. 2011, pp. 58–66.

[11] VU, T.-H., OLSSON, C., LAPTEV, I., OLIVA, A., AND SIVIC, J. Predicting actions from static scenes. In *ECCV, 421-436, 2014*.

[12] YAO, B., AND FEI-FEI, L. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *T-PAMI 34*, 9 (2012), 1691–1703.

[13] ZHA, Z.-J., ZHANG, H., WANG, M., LUAN, H., AND CHUA, T.-S. Detecting group activities with multi-camera context. *T-CSVT 23*, 5 (2013), 856–869.