

Human Activity Recognition from Frame's Spatiotemporal Representation

Zhipeng Zhao, Ahmed Elgammal
Computer Science Department, Rutgers University
110 Frelinghuysen Road, Piscataway, NJ 08854-8019, U.S.A
{zhipeng, elgammal}@cs.rutgers.edu

Abstract

This paper presents an approach for human activity recognition by representing the frames of the video sequence with the distribution of local motion features and their spatiotemporal arrangements. In this approach, the local motion features used for the representation of a frame are integrated from the ones detected in this frame and its temporal neighbors. The features' spatial arrangements are captured in a hierarchical spatial pyramid structure. By using frame by frame voting for the recognition, experiments have demonstrated improved performances over most of the other known methods on the popular benchmark data sets while approaching the best known results.

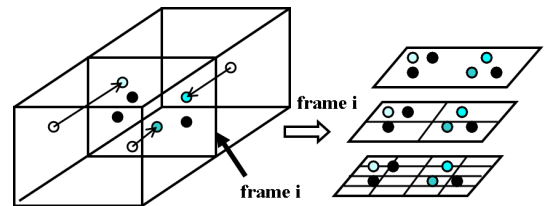


Figure 1. The representation for a frame i is built from the motion features detected in it and integrated from the nearby frames. The closer a features is to frame i , the higher weight it is assigned, represented by a darker circle. Then a spatial pyramid (e.g. $L = 2$) is applied to model the spatial arrangements among the features.

1. Introduction

Recognizing human activities from image sequences is an appealing yet challenging problem in computer vision with many applications including motion capture, human-computer interaction, environment control, and security surveillance. In this paper, we focus on recognizing the activities of a person in an image sequence from local motion features and their spatiotemporal arrangements.

Our approach is motivated by the recent success of “bag-of-words” model for general object recognition in computer vision [17, 11]. This representation, which is adapted from the text retrieval literature, models the object by the distribution of words from a fixed visual code book, which is usually obtained by vector quantization of local image visual features. However, this method discards the spatial and the temporal relations among the visual features, which could be helpful in the object recognition. Addressing this problem, our approach uses a hierarchical representation for the frames of the video sequence to integrate information from the spatial and the temporal domains. We first apply a spatiotemporal feature detector to the video sequence and obtain the local motion features. Then we generate a visual word code book by quantization of the local motion features and assign word label to each of them. Next for each frame, we integrate the visual words from its nearby frames, di-

vide the frame spatially into finer subdivisions and compute in each cell the histograms of the visual words detected in this frame and its temporal neighbors. Finally, we concatenate the histograms from all cells and use it as the feature for this frame. The representation for a frame i is illustrated in Figure 1. The contribution of our work lies in that besides the appearance information contained in the local motion features, our representation for the frame also captures both the spatial and the temporal relations among the features, which leads to better performance than the popular “bag-of-words” approach.

2 Related Work

Extensive research has been done in recognizing human activities. The approaches can be broadly categorized as model based, spatiotemporal template based and “bag-of-words” based. Model based approaches for activity recognition depend on locating and tracking body limbs in order to recognize the activity. That requires a model of the body, whether a 3D model or a 2D view-based model. We refer the reader to excellent surveys covering this topic, such as [2, 6]. However, for the task of activity recognition, tracking the limbs is not necessary. That motivates research on

obtaining spatiotemporal descriptors directly from the motion to recognize the activity without limb tracking. One of the earliest work on spatiotemporal descriptor was carried out by Polana and Nelson[12]. In Bobick and Davis’s work[3], Motion-Energy-Image and Motion-History-Image are introduced as templates for different motion recognition. Efros et al. [5] also proposed a spatiotemporal descriptor based on global optical flow measurements. Spatiotemporal template approaches are holistic approaches where global descriptors are used with no local features extracted.

In contrast, “bag-of-words” based approaches detect local salient descriptors as visual words, which are then used to recognize the activity. “bag-of-words” has been used successfully for object categorization[17, 11]. Inspired by text categorization, it represents the object as histogram of local features. Recently, “bag-of-words” methods have been used in activity recognition[14, 4, 15]. However, these approaches lack the relations between the features in the spatial and the temporal domains which are helpful for recognition. There are many recent research on extending “bag-of-words” to add the spatial relation in the context of object categorization [13, 1, 9, 7, 8]. In particular, pyramid match kernel [7, 8] used the weighted multi-resolution histogram intersection as a kernel function for classification with sets of image features.

The approach we propose here tries to simultaneously model the spatial and temporal relations of the local motion features. The temporal information is captured by integrating the local motion features from the temporally nearby frames and the spatial information is captured by using a hierarchical spatial pyramid in the representation.

3 Spatiotemporal Representation for the Frame

3.1 Feature extraction

We use the feature extractor from Dollar[4] for the local motion features’ detection and representation, which has been proven successful in [4, 10, 18]. In this method, the motion features are detected by applying separable linear filter to the video sequences. They are represented by the intensity gradients of a cuboid of spatiotemporally windowed data surrounding the detected interest point. To build the code book, we perform k-means from a random subset of motion features from the training data. The typical vocabulary size for our experiments is $K=250$.

3.2 Temporal integration of the motion features

Since a frame is correlated to its temporal neighbors, we build its representation from the motion features detected in it and its neighbor frames, weighed by the features’ temporal distance to this frame. Intuitively, the further the distance is, the less weight it should be assigned to. Therefore,

for a frame i , the weights assigned to the motion features from frame j are:

$$Weight(i, j) = e^{-\frac{dist(i, j)}{\sigma^2}} \quad (1)$$

where $dist(i, j)$ is the first norm distance between frame i and j and σ is the bandwidth for a smooth weight, which is empirically set to be 5 in our experiments. The temporal relations of the features to frame i are captured by the different weights. The weights are 1 for the motion features detected at frame i and are close to 0 for those from the distant frames. Therefore, only the motion features from nearby frames contribute significantly to the integration.

3.3 Spatial representation for the frame

With all the temporally weighted motion features for the frame, our next goal is to find a representation to model the spatial relations of these features in a way that it is suitable for measuring the similarity between the frames.

Let X and Y be two sets of motion features from two frames respectively. Inspired by [8], we represent the frame in a spatial pyramid. For each level $l, l = 0, \dots, L$, we divide the frame along x and y dimensions into $2^{2 \times l}$ subdivisions. Intuitively, we measure the distance between X and Y as the sum of the distances between the corresponding cells of all levels from X and Y . Each cell can be described as the histogram of the weighted motion features in it and the distances between them are measured by Chi-square distance. So the distance between X and Y is formulated as:

$$dist(X, Y) = \sum_{l=0}^L \sum_{i=1}^{2^{2 \times l}} \chi^2(H_X^l(i), H_Y^l(i)) \quad (2)$$

where $H_X^l(i)$ is the histogram of the weighted motion features from the i th cell in level l from X and $\chi^2(\cdot, \cdot)$ is the Chi-square distance. This is similar to representing X and Y by concatenating their histogram representations from all cells in all levels into a long histogram respectively and measuring their distance. Therefore, we can use these concatenated histograms as the representations for the frames.

Since different information is captured at various levels of the pyramid, different weights should be assigned to each of them. At finer resolution, the correspondence between two sets are captured more accurately. Therefore, we penalize the similarity information gained at a coarser level and give more weights to the similarity measured by the histogram distance at a finer resolution. The weight we assign at level l is: $weight(l) = 1/2^{L-l}$ for $l = 0, \dots, L$. The weighted distance between X and Y is:

$$dist(X, Y) = \sum_{l=0}^L \frac{1}{2^{L-l}} \times \sum_{i=1}^{2^{2 \times l}} \chi^2(H_X^l(i), H_Y^l(i)) \quad (3)$$

Because Chi-square distance satisfies

$$c\chi^2(a, b) = \chi^2(ca, cb) \quad (4)$$

where c is a scalar, we can directly embed the weight to the histogram representation. Putting everything together, our representation for a frame is the concatenated weighted histogram from all cells in all levels of the pyramid. In our representation, the temporal relations are modeled as the different weights assigned to the motion features and the spatial relations are captured in the spatial pyramid structure. With the motion features as the visual words, our representation simultaneously integrate appearance, spatial and temporal information.

Our representation for the frame is a straightforward extension of the popular “bag-of-words”. In each subdivision, all the local motion features are modeled as “bag-of-words”. When $L = 0$ and $\sigma = 0$, it reduces to the standard “bag-of-words” representation. In our experiments we observe that the performance does not improve much when $L > 1$. Therefore, we use the setting of $K = 250$ and $L = 1$, which leads to a 1250-dimension vector for the frame representation. For better computation efficiency, we normalize the vector by the total weight of all elements.

3.4 Recognition algorithm

Since the frames’ representations contain rich information in the spatiotemporal and the appearance domains, they can serve as classifiers for the underlying activity types. For each frame from the test frame, we label it with the closest frame in the training data sets and employ a majority voting throughout the sequence.

4 Experiments

4.1 Data sets

We carried out our experiments in three data sets, namely facial expressions data set from Dollar et al.[4], hand gestures data set from Wong et al.[18] and KTH human action data set from Schuldt et al.[14]. In all data sets, each video sequence contains one activity. The video sequences were converted into gray level to avoid the bias in color. The details of the data sets are summarized in Table 1 and some sample images from the video sequences are shown in Figure 2. In the experiments, we implemented a baseline approach using the “bag-of-words” representation.

4.2 Experimental results

With the same experimental setting on facial expression data set as in [4], we trained on one subject under one of the two lighting conditions and tested on: (1) the same subject under the same illumination, (2) the same subject under different illumination, (3) a different subject under the same

Dataset	Facial Expression	Hand Gesture	KTH
No. of classes	6	9	6
No. of subjects	2	2	25
No. of trials per subject	8	10	1
No. of conditions	2	5	4
Total No. of Samples	192	900	593

Table 1. Details of the data sets used in our experiments.

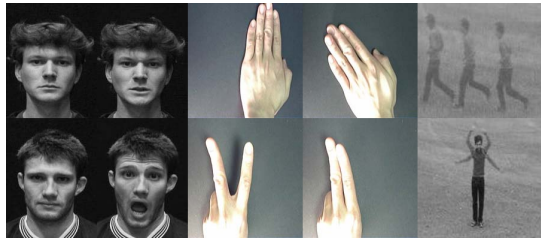


Figure 2. Sample images from the experiment data sets.

illumination, and (4) a different subject under different illumination. The recognition rates in each scenario from Dollar’s implementations[4], which is the baseline “bag-of-words” approach, and from our approach are shown in Table 2. In the first scenario, the recognition task is easy. The baseline algorithm already achieved very high recognition rate, therefore our approach only slightly improved the results. For the rest cases, our approach has demonstrated significant improvements.

With leave-one-out cross-validation experimental setting, we tested the baseline and our proposed methods on all data sets. The confusion matrices from our method are reported in Figure 3. The average recognition rates for all data sets, compared with other published results, are shown in Table 3. This has demonstrated that our approach can improve the “bag-of-words” baseline model and outperform most of the other known methods while approaching the best known result.

Methods	same sub. same illu.	same sub. diff. illu.	diff sub. same illu.	diff sub. diff illu.
Baseline	98.83	90.46	58.67	47.71
Our method	98.92	95.80	74.38	70.57

Table 2. The facial expression recognition rates(%)in different scenarios from the baseline “bag-of-words” algorithm and from our spatiotemporal representation.

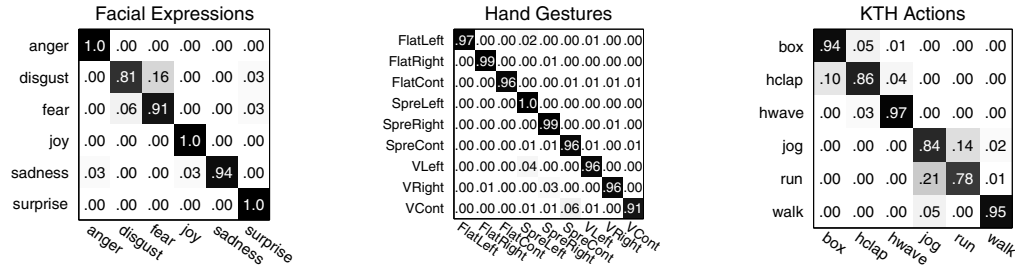


Figure 3. The Confusion matrices on all three data sets from our method

Methods:	Facial Expression	Hand Gesture	KTH
Baseline	91.33	85.81	81.51
Our method	94.33	96.78	89.67
Wong et al. [18]	83.33	91.47	83.92
Niebles et al. [10]	none	none	81.50
Wang et al.[16]	none	none	92.43

Table 3. The average recognition rates (%) for facial expression, hand gesture and KTH human action data sets obtained from different algorithms.

5 Discussion

We have presented a spatiotemporal frame representation for human activity recognition. Our approach simultaneously integrates the spatiotemporal relations among local motion features with their appearance information and embeds these rich information in the representation for the frames.

Our work differs from the pyramid match kernels[7, 8] in that: 1) Our goal is to find a suitable representation to integrate the spatiotemporal relations among motion features. The work in [7, 8] is seeking a suitable kernel function for two sets of image features. 2) Because our representation is a concatenated histogram, we measure the distance by Chi-square distance. The pyramid match kernels use histogram intersection as the distance function to satisfy the Mercer’s condition.

In the future, we intend to search systematically for the parameters in our approach and investigate methods for selecting informative key frames to speed up our current frame by frame voting method.

Acknowledgments: This research is partially funded by NSF CAREER award IIS-0546372.

References

[1] A. Agarwal and B. Triggs. Hyperfeatures: Multilevel local coding for visual recognition. In *ECCV06*, pages I: 30–43,

2006.

[2] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. *Comput. Vis. Image Underst.*, 73(3):428–440, 1999.

[3] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.

[4] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, October 2005.

[5] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV03*, page 726, 2003.

[6] D. M. Gavrila. The visual analysis of human movement: A survey. *Comput. Vis. Image Underst.*, 73(1):82–98, 1999.

[7] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV05*, pages 1458–1465, 2005.

[8] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR06*, pages 2169–2178, 2006.

[9] M. Marszaek and C. Schmid. Spatial weighting for bag-of-features. In *CVPR06*, pages 2118–2125, 2006.

[10] J. Niebles, H. Wang, and F. Li. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC06*, page III:1249, 2006.

[11] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR06*, pages 2161–2168, 2006.

[12] R. Polana and R. Nelson. Detecting activities. In *DARPA93*, pages 569–574, 1993.

[13] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlatons. In *CVPR06*, pages 2033–2040, 2006.

[14] C. Schudt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *ICPR04*, pages III: 32–36, 2004.

[15] C. Thureau. Behavior histograms for action recognition and human detection. In *HUM07*, pages 299–312, 2007.

[16] Y. Wang, P. Sabzmejdani, and G. Mori. Semi-latent dirichlet allocation: A hierarchical model for human action recognition. In *HUM07*, pages 240–254, 2007.

[17] J. Willamowski, D. Arregui, G. Csurka, C. R. Dance, and L. Fan. Categorization nine visual classes using local appearance descriptors. In *IWLAVS*, 2004.

[18] S. Wong, T. Kim, and R. Cipolla. Learning motion categories using both semantic and structural information. In *CVPR07*, pages 1–6, 2007.