# Unsupervised Action Classification Using Space-Time Link Analysis

Haowei Liu*, Rogerio Feris†, Volker Kruger‡, Ming-Ting Sun*
*University of Washington, Seattle, Washington 98195, USA, Email: hwliu, mts@u.washington.edu
†IBM T. J. Watson Research Center, Hawthorne, NY 10532, USA, Email: rsferis@us.ibm.com
‡CVMI, CIT, Aalborg University, Denmark, Email: vok@cvmi.aau.dk

*Abstract*—**In this paper we address the problem of unsupervised discovery of action classes in video data. Different from all existing methods thus far proposed for this task, we present a space-time link analysis approach which matches the performance of traditional unsupervised action categorization methods in a standard dataset. Our method is inspired by the recent success of link analysis techniques in the image domain. By applying these techniques in the space-time domain, we are able to naturally take into account the spatio-temporal relationships between the video features, while leveraging the power of graph matching for action classification. We present an experiment to demonstrate that our approach is capable of handling cluttered backgrounds, activities with subtle movements, and video data from moving cameras.**

## I. Introduction

How to automatically discover and recognize activities from video data is an important topic in computer vision. A solution to this problem will not only facilitate applications, such as video retrieval or summary, but will also improve, e.g., automatic video surveillance systems [1] and human-machine/robot communication [2]. In addition to its importance for many practical applications, unsupervised action categorization is important in the context of machine learning, particularly on how video processing approaches could allow a high-level ”understanding” of the data.

Numerous techniques have been proposed to solve the action classification problem [3]. The requirements of video analysis techniques are manifold, such as dealing with cluttered background, camera motion, occlusion, geometric and photometric variability, etc. [4], [5], [1]. Recently, unsupervised methods based on bag of visual words have become very popular as they could achieve excellent performance in standard datasets [6] and long surveillance videos [1], [7].

Generally, these unsupervised algorithms extract spatio-temporal feature descriptors called video words and then use document-topic models such as pLSA [8], LDA [9], or HDP [10] to discover latent topics [1], [7], [5]. A common limitation of these models is that they usually do not consider spatio-temporal correlations among visual words unless the correlations are represented explicitly [6]. Another general limitation is that some of these methods are EM-based learning approaches which makes recursive learning and updating difficult.

In this paper we introduce link analysis-based techniques to unsupervised activity discovery in video data that naturally preserves the spatio-temporal topology among the video words. Link analysis techniques are known from data mining, the information retrieval research communities, and the WWW [11]. They were largely ignored in computer vision until their recent introduction to the community by Kim et al. [12], who applied link analysis to unsupervised image clustering with impressive results.

The first step of our approach is to extract spatio-temporal features from the video data. Then, we construct a visual similarity network (VSN) [12] by computing the pairwise similarity between the features. Next, the VSN is analyzed by using the link analysis techniques, PageRank [11] and structure similarity (SS) [13], to produce an affinity matrix between all video sequences. Here, we interpret the pairwise matching weights as *votes* for the importance of the nodes which allows a quick division between consistent nodes and irrelevant ones (e.g., those from the background). Eventually, spectral clustering is applied to the affinity matrix to identify potential action categories. Link analysis techniques have been shown to be able to detect consistent matches (*hubs*) very effectively and efficiently [11], [12], [14], [15]. All computation and inference is done on the link weights between the nodes in the VSN which makes it fast and efficient. The key contributions of our work is that we extend link analysis techniques to the spatio-temporal domain and show that unsupervised discovery of action classes can greatly benefit from such approach and report results that match or exceed the performance of the state-of-the-art techniques in a standard dataset.

The paper is organized as follows: Section 2 describes our approach in detail, including the spatio-temporal interest point detector, the matching process, and link analysis techniques. In Section 3, we show the performance of our approach on a standard dataset and finally, Section 4 concludes our paper.

## II. Link-analysis for Spatio-Temporal Features

In this section, we break down our approach into its major components and give a detailed introduction to them. In detail, we will discuss the types of features we used, the use of shape context features, PageRank, structure similarity computation, and spectral clustering.

### A. Extraction of Spatio-Temporal Features

The first step of our action classification approach is to extract spatio-temporal interest points from the input video

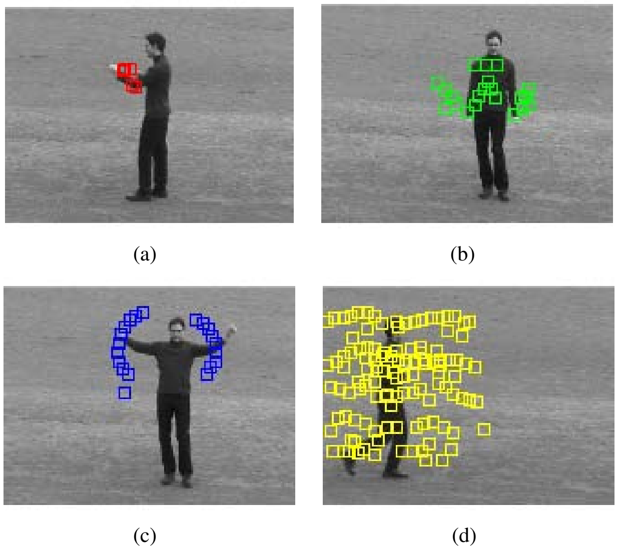

(a)
(b)
(c)
(d)

Fig. 1. Sample sequences with detected interest points for the KTH dataset. From (a) to (d), the activities are boxing, handclapping, hand-waving and jogging.

sequences. The two most recent spatio-temporal descriptors are proposed by Laptev and Lindeberg [16] and Dollar et al. [17] respectively.

We use the interest point detector proposed by Dollar et al. [17] in order to get denser spatio-temporal visual words. For a video sequence with pixel values $I(x, y, t)$, separable linear filters are applied to the video in order to obtain the response function as follows:

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \tag{1}$$

where $*$ indicates the convolution, $g(x, y, \sigma)$ is the 2D Gaussian smoothing kernel applied only along the spatial dimensions $(x, y)$, and $h_{ev}$ and $h_{od}$ are a quadrature pair of 1D Gabor filters applied temporally, which are defined as

$$h_{ev}(t; \tau, f) = \cos(2\pi f t)e^{-t^2/\tau^2} \tag{2}$$

$$h_{od}(t; \tau, f) = \sin(2\pi f t)e^{-t^2/\tau^2} \quad . \tag{3}$$

The two parameters $\sigma$ and $f$ correspond to the spatial and temporal scales of the detector respectively. The frequency of the harmonic functions is given by *f*. In all cases we use $f = 4/\tau$, as in [5].

Any region with spatially distinguishing characteristics undergoing a complex, non-translational motion induces a strong response [17]. At these interest points, we extract spatio-temporal volumes (cuboids). Later we calculate the brightness gradients within these volumes and concatenate them to form a feature vector. PCA is then used to reduce the dimensions of these feature vectors. Figure 1 shows the extracted interest points on a few sequences from the KTH dataset [18]. Considering Figure 1(c) as an example, we can see that the interest points occur at places around the arms, where the periodic motion induces strong responses.

### *B. Matching Spatial-Temporal Words and Building VSN*

Suppose we have a set $I$ of video sequences, each with $m_a, a \in I$, spatio-temporal features, and the total number of features in all sequences is $M$. We apply a graph matching algorithm on each pair of sequences $a, b \in I$ in order to determine feature level similarities. For computational efficiency reasons, we use the Hungarian method [19], a linear assignment matching approach, to match the extracted features. We also incorporated a shape context descriptor [20] to implicitly model the spatial arrangement of features.

Based on the pairwise matching results, and similarly to Kim et al. [12], we build a VSN $G = (V, E, W)$ where each node $a_i \in V$ represents the $ith$ feature in the input video $a$, and $b_j \in V$ represents the $jth$ feature in the input video $b$. The weights $w_e \in W$ for each edge $e = (a_i, b_j) \in E$ encode the similarity score between features $a_i$ and $b_j$. The similarity score between feature vector $a_i$ and $b_j$ is obtained through the exponential equation:

$$W(a_i, b_j) = exp(-cost(a_i, b_j)/2\sigma^2) \tag{4}$$

where $cost(a_i, b_j)$ is the matching cost between feature $a_i$ and $b_j$. In our experiments we compute the link weights simply from the difference between the two feature vectors with and without shape context features. For normalizing the weights we follow the approach outlined in [12].

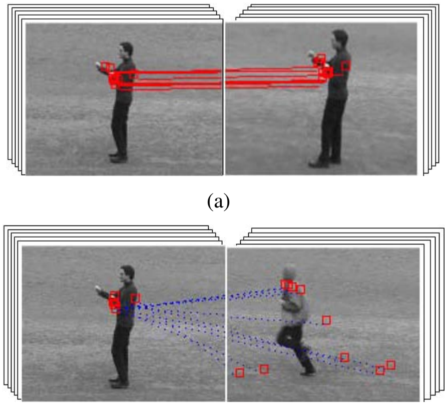
The intuition behind the matching algorithm and the VSN is that the number of links to and from a node reflects the co-occurrence statistics while each link weight reflects the belief in that match. This creates a clustering effect. The hope is that a) features from the same category would tend to interconnect with each other through strong links, while only weak links would exist between features from different categories, and b) features that appear often will have many links. Figure 2 shows the matching results between sequences from same and different categories respectively. As one can see, sequences from different classes would incur worse matching while the matching between sequences from the same category are more consistent and regular.

*1) PageRank:* The aim of the next step is to identify the strongest and most consistent features in each of the videos. We do this by extracting the sub-graph $G_a$ from our original VSN that contains the nodes from the video $a$ as well as all other nodes in the VSN that are connected to the nodes from $a$: we set $W_{ij} = 0$ if $i \notin a$ and $j \notin a$. Then, we apply pagerank [11] to the sub-graph $G_a$. The intuition behind the application of pagerank is that the nodes that are referenced (linked) often by important nodes are considered important as well. After pagerank, the features with high ranking values are those highly relevant and most consistent in the video $a$.

In short, the pagerank algorithm generates a pagerank vector P by solving the equation:

$$P = (1 - \alpha)(W + D)P + \alpha u \tag{5}$$

where $W$ is the weight matrix of $G_a$, $\alpha$ is a weighting constant set to 0.1 as in [12], $u$ is the transport vector



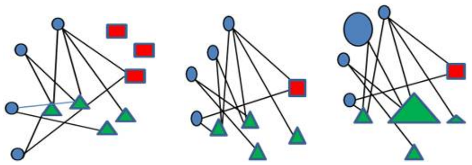
(a)

(b)

Fig. 2. The matching result between two sequences from (a) the same category and (b) different categories. Solid lines indicate matching pairs with low costs while dotted lines indicate costly matching pairs.

representing the initial prior of $P$ (set to a uniform distribution here) and $D = ud^T$, where d is the r-dimensional indicator vector identifying the nodes with zero outdegree and r is the dimension of the transport vector. The final ranking value of each node represents its relative importance in the VSN.

The process is illustrated in Figure 3. Initially, as Figure 3(a) shows, we have a VSN composed of features from three sequences. We extract the subgraph with respect to the first sequence, of which the features are represented as the circular nodes (Figure 3(b)). Then, we apply pagerank to the subgraph to determine the relative importance of the features in the subgraph. Figure 3(c) shows the final graph after pagerank. Larger nodes are those relevant features with respect to sequence one.



(a) (b) (c)

Fig. 3. The process of pagerank. (a) is the original similarity network we have. (b) shows the result after the subgraph extraction. Nodes of different shape represent features from different categories. After pagerank, features that are important would receive high ranking values, represented as the size of the nodes in (c). The larger a node is, the higher it ranks.

*2) Structure Similarity:* After computing pagerank, we evaluate the structure similarity [13] between two nodes. Here, we follow the reasoning in [13], [12]: Nodes with a similar set of links, i.e., nodes that are pointed to by a similar set of nodes and which are pointing to a similar set of nodes will most likely belong to the same category. We follow Brondel et al. [13] to compute the structure similarities $Z(a_i, b_j)$ between the visual word $a_i$ in sequence $a$ and $b_j$ in sequence $b$.

*3) Spectral Clustering:* By fusing the result of pagerank and structure similarity, we can obtain the similarity score between sequence $a$ and sequence $b$ by

$$S(a,b) = \sum_{b_j \in X_b} P_a(b_j) + \sum_{a_i \in X_a, b_j \in X_b} P_a(a_i) Z(a_i, b_j) \quad (6)$$

With the affinity matrix S at hand, we apply spectral clustering [21] on the k-nearest neighbor graph to uncover the underlying activities.

## III. Experiments

### A. KTH Dataset

In this section, we apply our algorithm to a standard dataset and show that it performs well compared to the state-of-the-art approaches.

The KTH dataset [18] is by far the largest standard activity dataset, which consists of six categories of activities including ”boxing”, ”hand clapping”, ”hand waving”, ”walking”, ”jogging” and ”running” performed by twenty-five actors in four different scenarios resulting in a total of 600 sequences. We test our algorithm using the same parameters as in [5]. The feature detector parameters are set to $\sigma = 2$ and $\tau = 2.5$. Each spatio-temporal patch is represented by the concatenated vector of its 3D gradients and then further reduced to 100 dimensions using PCA. We then apply our approach to cluster the video sequences. The confusion matrix for the KTH dataset is shown in table I. Note that we lump ”jogging” and ”running” into one category, as we did not incorporate features such as speed to distinguish these two activities. Our approach achieves 91.3% accuracy and performs well compared to the that of state-of-art approaches (e.g. Niebles et al. [5] also recently reported 91.3% considering running and jogging lumped together). The most confusing activities are ”boxing” and ”hand clapping”, both involving similar hand actions.

TABLE I
Confusion matrix for the KTH dataset. The average performance is 91.3%. ”box”, ”hc”, ”hw”, ”j/r”, ”walk” represent boxing, handclapping, handwaving, jogging/running, and walking respectively. For example, row one means out of all the boxing sequences, 84% are classified correctly and 16% are classified as handclapping.

| Category | box | hc | hw | jc | wa |
|---|---|---|---|---|---|
| box | ***0.84*** | 0.16 | 0 | 0 | 0 |
| hc | 0.04 | ***0.93*** | 0.03 | 0 | 0 |
| hw | 0 | 0.06 | ***0.94*** | 0 | 0 |
| jc | 0 | 0 | 0 | ***0.94*** | 0.06 |
| wa | 0 | 0.07 | 0 | 0.04 | ***0.9*** |

### B. Skating Dataset

As a second experiment, we apply our approach to a real world skating dataset reported in [22]. We extract 24 video sequences from the dataset and apply the same process to uncover three activities: stand-spin, sit-spin, and camel-spin. The detector parameters are set to $\sigma = 2$ and $\tau = 1.2$ when extracting the spatio-temporal interest points, which are then described by the corresponding PCA- reduced 3D gradients.

Figure 4(a) shows a frame for a sequence from the skating dataset with detected interest points. Since the sequences are shot with cluttered backgrounds and irregular camera motions, lots of irrelevant interest points are detected in the background. However, after space-time link analysis is applied, most of them are removed and not considered when classifying the sequences (Figure 4(b)). The average performance is 83.4%, which is better than 80.3% using the state of the art approach [5].

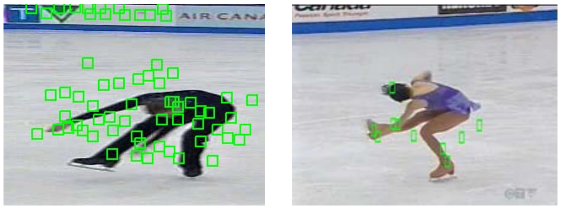

(a) (b)

Fig. 4. (a) Detected (noisy) interest points. (b) Highly ranked interest points after PageRank.

### C. Real World Surveillance Video

As a third experiment, we apply our approach to a real world surveillance system deployed in large retail stores to detect fraud scannings at the counters by differentiating three activities: pickup, scan, and drop. Figure 5 shows sample frames for these activities, with the detected interest points. We are able to achieve an 81.5% average accuracy.

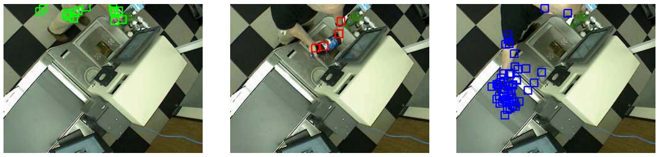

(a) (b) (c)

Fig. 5. Sample frames for three activities at the counter. Detected interest points are show in rectangles. (a), (b), (c) represents pickup, scan, and drop, respectively.

## IV. Conclusion

In this paper, we proposed a link-analysis based approach to unsupervised activity recognition. Different from previous approaches based on the bag of words models, the link-analysis approach takes into account the spatio-temporal relationship between visual words in the matching process. We see this as the major reason for the good performance of our approach.

Furthermore, we have tested the link-analysis on two datasets and a real world surveillance application, where our approach demonstrated its ability to deal with subtle movements, cluttered, and moving cameras. Future work will be to deal with multiple moving individuals/objects in the video data and more extensive evaluations. We would also like to evaluate the performance of our approach using better matching algorithm.

## References

[1] X. Wang, X. Ma, and E. Grimson, “Unsupervised activity perception by hierarchical bayesian models,” in *IEEE Conference on Computer Vision and Patter Recognition*, 2007.

[2] V. Krueger, D. Kragic, A. Ude, and C. Geib, “Meaning of action,” *Int. Journal on Advanced Robotics, Special issue on Imitative Robotics, T. Inamura and G. Metta (eds.)*, 2007.

[3] T. Moeslund, A. Hilton, and V. Krueger, “A survey of advances in vision-based human motion capture and analysis,” *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 90–127, 2006.

[4] J. C. Niebles and L. Fei-Fei, “A hierarchical model of shape and appearance for human action classification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[5] J. C. Niebles, H. Wang, and L. Fei-Fei, “Unsupervised learning of human action categories using spatial-temporal words,” *International Journal of Computer Vision.*, vol. 79(3), 2008.

[6] S. Savarese, A. D. Pozo, J. Niebles, and L. Fei-Fei, “Spatial-temporal correlations for unsupervised action classification,” in *IEEE Workshop on Motion and Video Computing*, 2008.

[7] X. Wang, K. T. Ma, G. W. Ng, and E. Grimson, “Trajectory analysis and semantic region modeling using a nonparametric bayesian model,” in *IEEE Conference on Computer Vision and Patter Recognition*, 2008.

[8] T. Hofmann, “Probabilistic latent semantic analysis,” in *Conference on Uncertainty in Artificial Intelligence*, 1999.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, January 2003.

[10] Y. Teh, M. Jordan, M. Beal, and D. Blei, “Hierarchical dirichlet process,” *Journal of the American Statistical Association*, 2006.

[11] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” in *Proceedings of the seventh international conference on World Wide Web*, vol. 7, 1998, pp. 107–117.

[12] G. Kim, C. Faloutsos, and M. Hebert, “Unsupervised modeling of object categories using link analysis techniques,” in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2008.

[13] V. D. Blondel, A. Gajardo, M. Heymans, P. Senellart, and P. V. Dooren, “A measure of similarity between graph vertices: Applications to synonym extraction and web searching,” *SIAM Review*, vol. 46(4), 2004.

[14] M. Najork and N. Craswell, “Efficient and effective link analysis with precomputed salsa maps,” in *Proceeding of the 17th ACM conference on Information and knowledge management*, Napa Valley, California, USA, 2008, pp. 52–63.

[15] M. Thelwall, *Link Analysis: An Information Science Approach*. Academic Press, 2004.

[16] I. Laptev and T. Lindeberg, “Space-time interest points,” in *IEEE International Conference on Computer Vision*, 2003.

[17] P. Dollar, V. Rabaud, G. Cottrellm, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” in *PETS*, 2005.

[18] C. Schuldt, I. Laptev, and B. Caputo, “Recognizing human actions: A local svm approach,” in *Proceedings of the International Conference on Pattern Recognition*, 2004.

[19] H. W. Kuhn, “The hungarian method for the assignment problem,” *Naval Research Logistics Quarterly*, vol. 2, 1955.

[20] S. Belongie, J. Malik, and J. Puzicha, “Shape matching and object recognition using shape contexts,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 24(4), 2002.

[21] Y. Song, W.-Y. Chen, H. Bai, C.-J. Lin, and E. Chang, “Parallel spectral clustering,” in *ECML*, 2008.

[22] Y. Wang, H. Jiang, M. S. Drew, Z. Li, and G. Mori, “Unsupervised discovery of action classes,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.