Action Recognition with Temporal Relationships

Guangchun Cheng Yiwen Wan Wasana Santiteerakul Shijun Tang Bill P Buckles
University of North Texas
{guangchuncheng, yiwenwan, wasanasantiteerakul, shijuntang}@my.unt.edu bbuckles@cse.unt.edu

Abstract

Action recognition is an important component in human-machine interactive systems and video analysis. Besides low-level actions, temporal relationships are also important for many actions, which are not fully studied for recognizing actions. We model the temporal structure of low-level actions based on dense trajectory groups. Trajectory groups are a higher level and more meaningful representation of actions than raw individual trajectories. Based on the temporal ordering of trajectory groups, we describe the temporal structure using Allen's temporal relations in a discriminative manner, and combine it with a generative model using bag-of-words. The simple idea behind the model is to extract mid-level features from domain-independent dense trajectories and classify the actions by exploring the temporal structure among them based on a set of Allen's relations. We compare the proposed approach with bag-of-words representation using public datasets, and the results show that our approach improves recognition accuracy.

1. Introduction

Action recognition is playing increasingly important roles in sports and human-computer interactive games. The major challenge for video-based applications is to recognize action or motion patterns from noisy and redundant visual information. The key issues involved include action representation and the modeling of spatial and temporal relationship between low-level actions.

Existing methods for vision-based action recognition can be classified into two main categories: feature-based bag-of-words and state-based model match. "Bag-of-words" has been successfully extended from text processing to many activity recognition tasks [4, 9]. Spatio-temporal relations between words in bag-of-words representation are not used. State-based matching methods establish a model to describe the temporal ordering of motion segments, which can discriminate between activities even for those with the same features with different temporal ordering. Methods in this category typically use hidden Markov models (HMMs) or spatio-temporal templates. Difficulties with model match methods include the determination of the model's structure and the parameters.

In this paper, a model combining temporal structure between features is proposed to explore the temporal relationships among the features for action recognition in videos. In order for a generic model that can be extended to different applications, dense trajectories are employed as observations, which are divided into meaningful groups by a graph-cut based aggregation method. A dictionary for these groups is learned from the training videos. In this study, we further explore the temporal relations between these trajectory groups. Each video is represented by combining bag-of-words and the temporal relationships between "words". We evaluate our model on public available datasets, and the experiments show that the performance is improved by combining temporal relationship and bag-of-words.

The contributions of this work are twofold. (1) In order to extend to different applications, our model uses groups of dense trajectories as its basis to represent actions in videos. Dense trajectories provide an effective treatment for cross-domain adaptivity [19, 14]. (2) The statistical temporal relationships among "words" is explored to improve the classification performance. The temporal relationships are intrinsic characteristics of actions and the connection between detected low-level action parts.

The remainder of this paper is organized as follows. After a brief introduction to related work, we describe the temporal structure in Section 3, and present how the learning is performed in Section 4. Section 5 gives experimental analysis with comparison with existing approaches. We conclude the work in Section 6.

2. Related Work

Here we briefly review studies that are closely related to this work.

Action recognition requires a discriminative description of the videos. Features such as trajectories and local descriptors are commonly used characteristics by encoding frequencies of spatial and/or temporal features. Trajectories extracted through tracking are widely used as ob-

servations to construct the codebook of "visual words". Many approaches encode the trajectories using a series of interest-point based descriptors such as 3D scale-invariant feature transform (SIFT) [16], 3D histogram of gradients (HoG) [7], histogram of optical flow (HoF), motion boundary histograms(MBH) [5], or combination of them. As pointed out by Wang [18], sparse interest points performed worse than dense sampling of tracking points for both image classification and action recognition. Based on this observation, Wang *et al.* [19] proposed an approach to describe videos by dense trajectories, and designed a descriptor to encode the dense trajectories for action recognition. While dense trajectories provide comprehensive information about the motion in the video, they are redundant and low-level representation to form meaningful codewords. As Liu *et al.* [10] stated, meaningful grouping of vision features within the original bag-of-words assists the classification. This notion inspires the basis of this paper.

The large number of dense trajectories makes it possible to perform statistical learning of the meaningful clusters. Lan *et al.* [8] and Raptis *et al.* [14] proposed to use action parts for action recognition and localization. Both models utilized latent variables and trained the models discriminatively. Lan *et al.* [8] constructed a chain-structured graph to represent the relations between features which are action bounding boxes. Spatial relations and temporal smoothness was used to construct the model, and the recognition was achieved by measuring the compatibility between a given video and the configurations of bounding boxes of actions with known labels. Raptis *et al.* [14] extracted mid-level action parts to express salient spatio-temporal structures of actions in videos, and constructed a graphical model to incorporate appearance and motion constraints for action parts and their dependencies. The action parts in [14] were obtained by forming clusters of trajectories, which is similar to what we use in this paper. However, Raptis *et al.* [14] didn't explore the temporal relations among the action parts. This paper develops a method to explore their dependencies and temporal constraints of action parts.

Most actions, especially high-level actions, are recognized depend on two components: meaningful short-term subactions (referred to as *actionlets* hereafter) and the spatial/temporal arrangement of them. The actionlets can be raw trajectories of tracked points, or a cluster of spatio-temporally similar trajectories [14] as stated before. Bag-of-words representation models the actionlets without explicit treatment for spatial/temporal relations. The spatial/temporal relations of actionlets are described by probabilistic models such as hidden Markov models [20] and dynamic Bayesian networks [6]. Unfortunately, these approaches generally assume fixed number of actions, and require large training sets in order to learn the model structure and their parameters. Bobick and Davis [3] described motion energy image and motion history image to represent the space-time volume of a specific action, and applied template matching for recognition. Description-based models incorporate expert domain knowledge into the definition of actions, and simplify the recognition in structured scenarios [1]. In order to express the temporal relationships, Allen [2] described 13 predicates to describe the temporal relations between any two time intervals. Many approaches are proposed using Allen's temporal predicates to express temporal relationships between actionlets [11, 15]. Most of such approaches are based on a logic description of the actions.

As observed from the aforementioned research, action recognition has attracted study from both feature-based and description-based approaches. The former is usually used as the basis for the latter, and the later draws closer to a human's understanding of an action. This paper recognizes actions by extracting mid-level actionlets which are represented by trajectory groups and exploring their temporal relations quantitatively using Allen's interval relations. These actionlets and their temporal relations are more expressive and can be integrated into other higher-level inference systems.

3. Temporal Structure of Trajectory Groups

In order to develop a application-independent approach for action recognition, we extract features to express meaningful *actionlets* based on dense trajectories. For raw trajectory descriptors, we employ the form that Wang *et al.* proposed [19]. There exists a mismatching gap between these raw trajectories and the description of common understanding of actions which are described categorically. In this paper, we therefore cluster these dense trajectories to meaningful groups, and construct a bag-of-words representation to describe these trajectory groups in Section 3.1 and Section 3.2. The temporal relations between "words" are described in Section 3.3.

3.1. Grouping dense trajectories

The dense trajectories in [19] are extracted from multiple spatial scales. The feature points are sampled on a grid basis, and the tracking of them is based on dense optical field. Abrupt change and stationary trajectories are removed from the final results. For each trajectory, the descriptor combines trajectory shape, appearance (HoG), and motion (HoF and MBH) information. Therefore, the feature vector for a single trajectory is in the form of

$$T = (S, HoG, HoF, MBH_x, MBH_y) \qquad (1)$$

where $S = \frac{(\Delta P_t, \ldots, \Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1} \|\Delta P_j\|}$ is the normalized shape vector, and its dimension L is the length of the trajectory. MBH

is divided into MBH_x and MBH_y to describe the motion in x and y direction respectively.

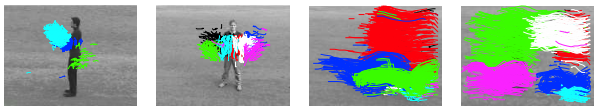
The trajectories are clustered into groups based on their descriptors, and each group consists of spatio-temporally similar trajectories which characterize the motion of a particular object or its part. We develop a distance metric between trajectories taking into consideration their spatial and temporal relations. Given two trajectories t_1 and t_2, the distance between them is

$$d(t_1, t_2) = \frac{1}{L} d_S(t_1, t_2) \cdot \bar{d}_{spatial}(t_1, t_2) \cdot d_t(t_1, t_2) \quad (2)$$

where d_S is the Euclidean distance between the shape vectors of t_1 and t_2, $\bar{d}_{spatial}(t_1, t_2)$ is the mean spatial distance between corresponding trajectory points, and $d_t(t_1, t_2)$ indicates the temporal distance. For simplicity, we use the following in experiments.

$$d_t(t_1, t_2) = \begin{cases} 1 & \text{TimeDiff}(t_1, t_2) < L \\ \infty & \text{otherwise} \end{cases} \quad (3)$$

Trajectories are grouped based on a graph clustering algorithm GANC [17]. As input to GANC, we compute an $n \times n$ affinity matrix A of the trajectories, with each element $a(t_i, t_j) = \exp^{-d(t_i, t_j)}$, where n is the number of trajectories in a video. GANC produces clusters minimizing the normalized cut criterion in a greedy agglomerative hierarchical manner. Figure 1 shows examples of grouped trajectories for some video samples.



(a) Boxing (b) Clapping (c) Running (d) Jogging

Figure 1. Illustration on trajectory grouping based on spatio-temporal proximity.

3.2. Bag of groups

The trajectories provide low level description to the action content in a video. A mean feature vector, $\overline{T_i}$, is obtained for all the trajectories in the same group. Because of the large motion variation in even the same type of actions, our model constructs a codebook for these trajectory groups, and assigns each group to its closest word in the codebook. The size of the codebook, D, is determined based on the experiments, and is set to 1000 in the experiments. K-means is used over $\overline{T_i}$'s to generate the words with Euclidean distance metric. In the following, $f : g \rightarrow c$ is used to indicate the mapping from a group to a group word.

Given the codebook, the trajectory groups of a video are assigned different words, and the video can have a bag-of-groups representation as follows, where d_i is the frequency of word c_i in the video.

$$BoG = \{d_1, d_2, ..., d_D\} \quad (4)$$

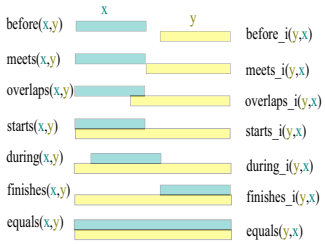


Figure 2. Allen's temporal relations between two intervals.

3.3. Temporal structure

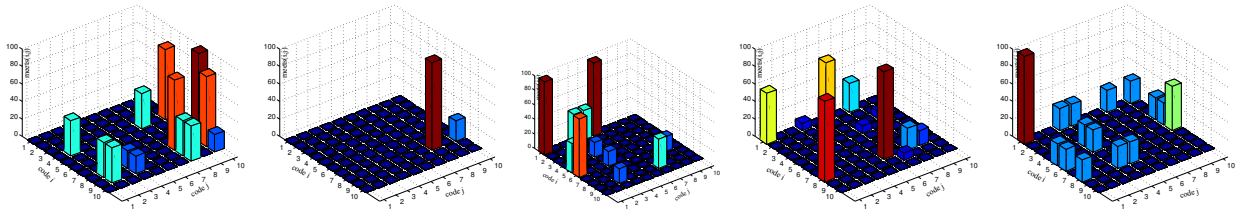
Based on the bag-of-groups representation, our model develops the statistical temporal relations between the "groups". According to Allen's conclusion, there exist thirteen temporal relations between two actions based on the actions' durance intervals, i.e. *before*($\mathcal{B}$), *meets*($\mathcal{M}$), *overlaps*($\mathcal{O}$), *starts*($\mathcal{S}$), *during*($\mathcal{D}$), *finishes*($\mathcal{F}$), *equals*($\mathcal{E}$), and their inverse relationships. *before_i* means *before inversely* (i.e. *after*), and the same for other relations on the right column. See Figure 2 for a summary on the temporal relations. We use seven temporal structures to express these relationships based on the symmetry existing among them as also noticed by Patridis *et al.* [13].

For each type of action, the temporal relationship between pairs of group words is modeled by seven two-dimensional histograms of pairwise relation between them. Each histogram shows the frequencies with which the relation is true between a pair of group words. That is, a temporal relation $\mathcal{R}_i \in \{\mathcal{B}, \mathcal{M}, \mathcal{O}, \mathcal{S}, \mathcal{D}, \mathcal{F}, \mathcal{E}\}$, $\mathcal{R}_i(x, y)$ is the frequency of $x\, \mathcal{R}_i\, y$ between two group words x and y. In our model, we construct the temporal relations for each type of action in a supervised manner, i.e. we learn discriminatively $p(\mathcal{R}_i|\alpha)$ for each action type α. Figure 3 shows an example of *meets* for different actions in one testing dataset. It can be observed that different actions exhibit different histograms, and similar actions have similar histograms. Examining each of the histograms shows which temporal relation (such as *meets* for boxing) has a stronger response for some pairs of group words than the others. This implies the major relation relations between *actionlets*.

We obtain the signature for action α by combining the bag-of-words and the temporal relations: $A = \{BoG^\alpha, \{\mathcal{R}_i^\alpha\}_{i=1}^7\}$, and this is used as the feature of the model.

During recognition a similar process is followed to extract the feature for the target video. Suppose it is F: $\{bog, \{\mathbb{R}_i\}_{i=1}^7\}$. We seek for action α^* which maximizes the likelihood:

$$\begin{aligned} \alpha^* &= \arg\max_\alpha \mathcal{L}(F|\alpha) \\ &= \arg\max_\alpha \prod \mathcal{L}(c_j|\alpha) \prod \mathcal{L}(\mathbb{R}_i|\alpha) \end{aligned} \quad (5)$$



(a) boxing (b) handclapping (c) jogging (d) running (e) walking

Figure 3. Histograms of temporal relation *meets* for five different actions in KTH dataset. The X and Y axes are the types of group codes, and the Z values are the frequency before normalization. Among them, histograms of jogging and walking are relatively close to each other. So are boxing and handclapping.

based on the assumption that different groups and temporal relations are independent.

$\mathcal{L}(c_j|\alpha)$ can be directly retrieved from the signature of action α, denoted as $p(c_j|\alpha)$ (see next section), and here we discuss how to obtain the likelihood of $\{\mathbb{R}_i\}_{i=1}^{7}$. We use the distance between $\mathcal{R}_i^\alpha$ and $\mathbb{R}_i$ to define the likelihood. Both $\mathcal{R}_i^\alpha$ and $\mathbb{R}_i$ are matrices, and their distance is defined according to [12] as follows in equation (6) where λ_j's are eigenvalues of matrix $\mathcal{R}_i^{\alpha\,-\frac{1}{2}}\mathbb{R}_i\mathcal{R}_i^{\alpha\,-\frac{1}{2}}$. In case when $\mathcal{R}_i^\alpha$ is singular, its pesudo-inverse is used as its inverse.

$$d(\mathcal{R}_i^\alpha, \mathbb{R}_i) = \sqrt{\sum_{j=1}^{L=2} (\log \lambda_j)^2} \tag{6}$$

The likelihood of $\mathbb{R}_i$ being action α is defined as $\mathcal{L}(\mathbb{R}_i|\alpha) = \frac{e^{-d(\mathcal{R}_i^\alpha, \mathbb{R}_i)}}{\int_{\mathbb{R}_i} e^{-d(\mathcal{R}_i^\alpha, \mathbb{R}_i)}\, \mathrm{d}\mathbb{R}_i}$. Disregarding the constant multiplier in nominator, and substituting into the equation (5) results in

$$\alpha^* = \arg\max_\alpha \; exp\left\{-\sum_{i=1}^{7} d(\mathcal{R}_i^\alpha, \mathbb{R}_i)\right\} \prod_{j=1}^{|W|} p(c_j|\alpha) \tag{7}$$

where $|W|$ is the number of group words in the video. This problem can be solved effectively when the signatures of known actions and the features of the target video are available. the solution is described in next section.

4. Learning and Recognition

To construct the signatures of actions, a supervised discriminative learning approach is applied to obtain the probability of every code given the action $p(c_i|\alpha)$ and the seven histograms for temporal relations. We learn the $p(c_i|\alpha)$ and the temporal histograms for each type of actions.

For a specific dataset, we assume that the actions α's are known, and the vocabulary should be learnt from it first. To obtain the bag-of-groups representation as described in Section 3.2, we combine and cluster the groups from all the videos.

We take simple methods to learn the conditional probability and the temporal histograms. Following a Bayesian training procedure, we count the occurrence (T_{c_i}) of each group word for all the videos with the same action, and then compute the conditional distribution using each word's frequency. The temporal histograms are constructed computed based on each video and are averaged over all videos. For each trajectory group in a video of action α, we compute its temporal distances to all of the other groups in that video, determine the Allen temporal relations between them, and count the frequency of each relation. The seven temporal histograms are updated correspondingly.

For recognition, the bag-of-groups and temporal histograms are extracted from each test video, and compared with learned action signatures based on the distance metric discussed in Section 3.3. The final decision is made using (7).

5. Experimental Results

In this section, we demonstrate experiments to evaluate our approach using the KTH human motion dataset and Weizmann action dataset. The actions in these two datasets were recorded in constrained settings. For simplicity, the experiments only use bag-of-words and a simple classifier for comparison. The experimental results show that the recognition accuracy improves by combining temporal information.

5.1. KTH dataset

The KTH dataset contains six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed several times by 25 subjects in four different scenarios, including outdoors, outdoors with scale variation, outdoors with different clothes, and indoors. All video sequences have static and homogeneous backgrounds. Altogether there are 2391 sequences.

We segmented a video sequence into clips of around 20 seconds if the video contains cluttered motion. This preprocessing reduces the number of the trajectories in a video for analysis, and does not affect the application of online

Table 1. Accurary for KTH dataset

	BoG	BoG+Temporal
boxing	96.0%	100.0%
handclapping	78.0%	84.0%
handwaving	88.0%	92.0%
jogging	70.0%	76.0%
running	98.0%	100.0%
walking	82.0%	86.0%
Mean Accuracy	85.3%	89.7%

action detection. For each category, we have 50 videos for training and 50 videos for testing. The average per-class classification accuracy are summarized in Table 1. The result for BoG is from using only bag-of-groups based on our implementation using a naive Bayesian classifier. Our model achieves 89.7% of accuracy by combining bag-of-groups and temporal relations. We can see the performance improvement compared with the result of BoG.

Table 2. Accuracy for Weizmann dataset

9-class Weizmann dataset

	BoG	BoG+Temporal
Mean Accuracy	90.2%	94.1%

10-class Weizmann dataset

	BoG	BoG+Temporal
Mean Accuracy	85.3%	87.8%

5.2. Weizmann dataset

This updated Weizmann dataset consists of 90 low-resolution video sequences showing nine different people, each performing 10 natural actions: bending, jumping-jack, jumping-in-place, running, gallop sideways, skipping, walking, waving one hand and waving two hands. 9 actions (not including skipping) were also used for experiments by researchers. The recognition results for both 10-action and 9-action are shown in Table 2.

6. Conclusion

We proposed an algorithm to explore the temporal relations between trajectory groups in videos, and applied it to action recognition and intelligent human-machine interaction systems. The trajectory groups are application-independent features, and work as mid-level descrptions of actions in videos. The experiments demonstrated its performance improvements compared with pure bag-of-features method. The success of this semantics-free recognition method provides the potential to define high-level actions using low-level actionlets and their temporal ordering. This is similar to the way humans perceive and recognize actions. The information extracted from the temporal relation between trajectory groups can be input to other inference engines.

References

[1] J. Aggarwal and M. Ryoo. Human activity analysis: a survey. *ACM Computing Surveys*, 43(3):1–43, 2011. 2

[2] J. F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983. 2

[3] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *PAMI*, 23(3):257–267, 2001. 2

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2

[6] D. Damen and D. Hogg. Recognizing linked events: Searching the space of feasible explanations. In *CVPR*, 2009. 2

[7] A. Klaser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In *BMVC*, 2008. 2

[8] T. Lan, Y. Wang, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In *ICCV*, 2011. 2

[9] I. Laptev, B. Caputo, C. Schuldt, and T. Lindeberg. Local velocity-adapted motion events for spatio-temporal recognition. *Computer Vision and Image Understanding*, 108(3):207–229, 2007. 1

[10] J. Liu, S. Ali, and M. Shah. Recognizing human actions using multiple features. In *CVPR*, 2008. 2

[11] V. I. Morariu and L. S. Davis. Multi-agent event recognition in structured scenarios. In *CVPR*, 2011. 2

[12] X. Pennec, P. Fillard, and N. Ayache. A Riemannian framework for tensor computing. *IJCV*, 66(1):41–66, 2006. 4

[13] S. Petridis, G. Paliouras, and S. J. Perantonis. Allen's hourglass: Probabilistic treatment of interval relations. In *7th International Symposium on Temporal Representation and Reasoning*, 2010. 3

[14] M. Raptis, I. Kokkinos, and S. Soatto. Discovering discriminative action parts from mid-level video representations. In *CVPR*, 2012. 1, 2

[15] M. Ryoo and J. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *ICCV*, 2009. 2

[16] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *International Conference on Multimedia*, 2007. 2

[17] S. S. Tabatabaei, M. Coates, and M. Rabbat. GANC: Greedy agglomerative normalized cut for graph clustering. *Pattern Recognition*, 45(2):831–843, Feb. 2012. 3

[18] H. Wang. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009. 2

[19] H. Wang, A. Klaser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *CVPR*, 2011. 1, 2

[20] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan. Modeling individual and group actions in meetings with layered HMMs. *IEEE Transactions on Multimedia*, 8(3):509–520, 2006. 2