LEARNING DIRECTIONAL CO-OCCURRENCE FOR HUMAN ACTION CLASSIFICATION

Hong Liu, Mengyuan Liu, Qianru Sun

Key Laboratory of Machine Perception, Peking University, China
E-mail: {hongliu, liumengyuan}@pku.edu.cn; qianrusun@sz.pku.edu.cn

ABSTRACT

Spatio-temporal interest point (STIP) based methods have shown promising results for human action classification. However, state-of-art works typically utilize bag-of-visual words (BoVW), which focuses on the statistical distribution of features but ignores their inherent structural relationships. To solve this problem, a descriptor, namely directional pairwise feature (DPF), is proposed to encode the mutual direction information between pairwise words, aiming at adding more spatial discriminant to BoVW. Firstly, STIP features are extracted and classified into a set of labeled words. Then in each frame, the DPF is constructed for every pair of words with different labels, according to their assigned directional vector. Finally, DPFs are quantized to be a probability histogram as a representation of human action. The proposed method is evaluated on two challenging datasets, Rochester and UT-interaction, and the results based on chi-squared kernel SVM classifiers confirm that our method can classify human actions with high accuracies.

Index Terms— Spatio-temporal interest point, bag-of-visual words, co-occurrence

1. INTRODUCTION

Human action classification is significant for smart surveillance, content-based video retrieval and human robot interaction, while it is still challenging due to clustered background, occlusion and other difficulties in video analysis. What's more, inter-similarity between different actions also brings serious ambiguity. Recently, several spatio-temporal interest points (STIPs) based works have shown promising results for describing actions [1][2][3]. These works first extract STIPs from training videos and cluster them into visual words using clustering methods. The bag-of-visual words (BoVW) model is then adopted to represent original action by a histogram of words distribution, and to train classifiers for classification. However, BoVW ignores the spatio-temporal distribution information among words and thus leads to misclassification for actions sharing similar words distribution. To make up for the above problem of BoVW, the spatio-temporal distribution of words is explored. Niebles *et al.* [4] used latent topic models such as the probabilistic Latent Semantic Analysis (pLSA) model to learn the probability distributions of words. And Ryoo [5] represented an action as an integral histogram of spatio-temporal words which models how words distributions change over time. Besides directly modeling the distributions, Burghouts *et al.* [6] brought in a novel spatio-temporal layout of actions which assigns a weight to each word by its spatio-temporal probability. These efforts attempted to encode spatio-temporal information using words in groups. Meanwhile, considering words in pairs is an effective alternative to describe the distribution of words.

Relation to prior work: Ryoo and Aggarwal [7] introduced a spatio-temporal relationship matching method which explored temporal relationships (e.g. before and during) as well as spatial relationships (e.g. near and far) among pairwise words. Savarese *et al.* [8] focused on the co-occurrence of pairwise words and proposed the usage of spatial-temporal correlograms which capture the co-occurrences in local spatio-temporal regions. To involve global relationships, we previously proposed to encode the co-occurrence correlograms by computing pairwise normalized google-like distances (NGLD) in [9]. Further, more temporal correlation among local words was added in [10]. These works show that co-occurrence pairs can properly represent the spatial information in the whole word set. In this work, we observe that human actions make huge senses in moving body parts directionally from one place to another. This phenomenon reflects the importance of directional information for action representation. Hence the attribute of mutual directions are assigned to pairwise points to encode additional structural information. Comparing with [7], our novelty lies in the use of direction instead of distance when describing the pairwise co-occurrence. Our work also differs from [8] and [9] in the use of both number and direction of pairwise words . A dimension reduction method is additionally introduced to form a rich representation with low dimension.

The rest of the paper is as follows: Sec.2 illustrates our new representation for action classification. Sec.3 discusses the experiments comparing with BoVW based methods and state-of-the-arts. Finally, conclusions are drawn in Sec.4.

This work is supported by National Natural Science Foundation of China(NSFC, No.60875050, 60675025), National High Technology Research and Development Program of China(863 Program, No.2006AA04Z247), Scientific and Technical Innovation Commission of Shenzhen Municipality (No.201005280682A, No.JCYJ20120614152234873, CXC201104210010A)

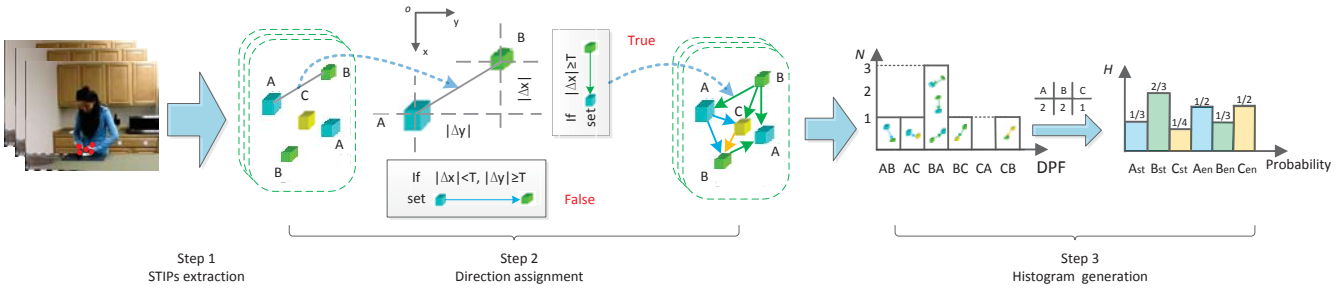


Fig. 1. Flowchart of extracting action representation. STIPs are detected and clustered into *K* words. For points in pairs, DPFs are constructed using their labels and directions (the criterion of direction assignment is in Step 2). Combining histogram of DPFs (named *N*) which has *K·(K-1)* bins with distribution of words, histogram H with *K·2* bins is formed as final representation.

2. PROPOSED METHOD

Pairwise relationships between STIPs were modeled in spatial domain resulting in a Co-occurrence Map [9] which lacks structural information to classify similar actions. This work presents a new representation to encode additional directional information between STIPs. As shown in Fig. 1, we first assign directions to proper pairwise words. For any pair, a directional pairwise feature (DPF) is then defined according to pairwise words' labels and assigned direction. Final representation is based on the statistics of DPFs.

2.1. DPF and Histogram of DPFs

The criterion of direction assignment is introduced before defining DPF. Using STIPs detection and clustering methods, a video is represented by a set of words with different labels. We just consider word pairs appearing in same frame with different labels. Sketch in step 2 (Fig.1) shows how to assign direction for A and B. Although the vector formed by A and B provides exact spatial information, it is not directly used as feature taking robustness into consideration. Instead, we only care whether the direction is from A to B or B to A. Vertical or horizontal relationship is utilized to figure out the direction between A and B with two reference directions defined from up to down and left to right respectively. We observe that human actions like waving right hand and waving left hand are usually symmetric. Their directional information are opposite in horizontal direction while the same in vertical direction. Thus, we consider the vertical relationship priority to the horizontal one to eliminate the ambiguity brought by symmetric actions. Let Δx and Δy represent projector distances and T stand for a threshold value. If A and B are far in vertical direction ($\Delta x \geq T$), the reference direction is set from up to down. In contrast ($\Delta x < T$), the relationship in the vertical direction is not stable and thus discarded. In this case, the horizontal relationship is checked similarly. As for A and B, since $\Delta x \geq T$ and B is on the top of A, the vertical relationship is selected and the direction is assigned from B to A, which is in accordance with the reference direction.

We are now ready to define DPF and histogram of DPFs. Suppose words are clustered into K labels for a given video. $S = \{S_1, ..., S_k, ..., S_K\}$ denotes the word set and S_k contains all words labeled $k \in \{1, ..., K\}$. $pt_i = (x_{pt_i}, y_{pt_i}, t_{pt_i})$ represents a word labeled i appearing in frame t_{pt_i}. x_{pt_i} and y_{pt_i} are the horizontal and vertical coordinates values. If a direction is assigned from i to j, a $\text{DPF}_{i,j}$ is defined. $n(pt_i, pt_j)$ in Formula (1) is utilized to record whether there exists $\text{DPF}_{i,j}$ between pt_i and pt_j.

$$n(pt_i, pt_j) = \begin{cases} 1, & if\ (t_{pt_i} = t_{pt_j}) \wedge \\ & \quad \{(|\Delta x| < T, |\Delta y| \geq T, y_i < y_j) \\ & \quad \vee\ (|\Delta x| \geq T, x_i < x_j)\}, \\ 0, & otherwise \end{cases} \tag{1}$$

Threshold value T in Formula (1) is related to the spatial scope of whole words and we use the average distance between pairwise words in Formula (2) to represent T.

$$T = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{\forall pt_i \in S_i, \forall pt_j \in S_j} |x_{pt_i} - x_{pt_j}|}{\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{\forall pt_i \in S_i, \forall pt_j \in S_j} 1} \tag{2}$$

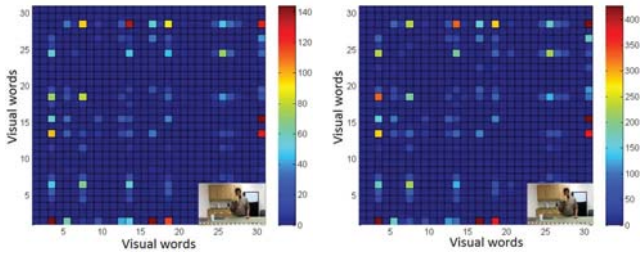
Let DPFs denotes $\{\text{DPF}_{i,j}\}$ for $(i, j \in \{1, ..., K\}, i \neq j)$. Formula (3) calculates the number of $\text{DPF}_{i,j}$ named $N(i, j)$ for all word pairs in S. And *N* is the histogram of DPFs:

$$N(i,j) = \sum_{\forall pt_i \in S_i, \forall pt_j \in S_j} n(pt_i, pt_j) \tag{3}$$

The extracted histogram of DPFs is most related to the Co-occurrence Map which records the number of co-occurrence between STIPs labeled i and j for position $(i, j)(i, j \in (1, ..., K))$. In order to directly show the difference, an action "eating a banana" is utilized. To facilitate observation, 800 STIPs are extracted from the action and clustered to 30 labels. Two result maps: histogram of DPFs (displayed in matrix form) and Co-occurrence Map are shown in Fig. 2. We note that element values in (i, j) and (j, i) are the same in Co-occurrence Map while different in the histogram of DPFs. Thus, the histogram of DPFs encodes more distinct information than Co-occurrence Map.

2.2. Representation Generation

After computing histogram of DPFs, a video is reduced to a histogram N with *K·(K-1)* dimension which is still high. Further dimension reduction is needed for realtime applications. Any word labeled i is short for i below. In Formula



(a) Histogram of DPFs (asymmetry) (b) Co-occurrence Map (symmetry)

Fig. 2. The Histogram of DPFs (in matrix form) (a) and Co-occurrence Map (b) are respectively extracted for "eating a banana" in Rochester.

(4), $N(i,j)$ refers to the number of $\text{DPF}_{i,j}$ and $N(i)$ is the number of i. $P(DPF_i^{st}|DPFs)$ represents the probability of appearing i as a start point given DPFs:

$$P(DPF_i^{st}|DPFs) = \frac{\sum_{j=1}^{K} N(i,j)}{\sum_{j=1}^{K}\{N(i)\cdot N(j)\}} \quad (4)$$

Similarly, $P(DPF_i^{en}|DPFs)$ in Formula (5) represents the probability of appearing i as an end point:

$$P(DPF_i^{en}|DPFs) = \frac{\sum_{j=1}^{K} N(j,i)}{\sum_{j=1}^{K}\{N(i)\cdot N(j)\}} \quad (5)$$

Above two probability values are combined in Formula (6) to construct final representation H whose dimension is $K \cdot 2$. Given two similar actions sharing *i* and *j*, H should be similar despite the difference of the whole number of *i* or *j*. Using H instead of original histogram N, the compression ratio of dimension equals $(K-1)/2$:

$$H = \{\{P(DPF_i^{st}|DPFs)\}_{i=1}^{K}, \{P(DPF_i^{en}|DPFs)\}_{i=1}^{K}\} \quad (6)$$

Histogram of DPFs and H is illustrated in histogram generation step of Fig. 1. A_{st} and A_{en} in histogram H means the probability of A appearing as a start point and an end point respectively. Noting that A_{st} plus A_{en} is no more than 1, since the relationships between some pairs are discarded taking (C,A) in the sketch of Fig. 1 as an example. If relationships between A and all other points are considered, the value A_{st} plus A_{en} should equal 1.

2.3. Representation Extraction Algorithm

To extract action representation which is named H from video $\{I_t\}_{t=1}^{F}$ with F frames, the procedure is as follows: (I) STIPs $S = \{(x,y,t) \mid (x,y) \in I_t, 1 \le t \le F\}$ are detected and represented as $\{des_{(x,y,t)}\}$ by local patches around STIPs. (II) $\{des_{(x,y,t)}\}$ are clustered into K centers. Let $N(k)$ records the number of STIPs labeled $k(1 \le k \le K)$ and labeled points are stored in $S = \{S_1, S_2, ..., S_K\}$. (III) Get T as a threshold value from Formula (2) for the direction assignment section, and use Formula (1),(3) to calculate $N(i,j)$ for $(1 \le i,j \le K, i \ne j)$. After this step, histogram N with $K \cdot (K-1)$ dimension is obtained. (IV) N is finally compressed as H with $K \cdot 2$ bins using Formula (4-6). All steps are shown in Algorithm 1. The algorithm focus on pairwise features and extracting directional information from them to reflect the natural structure of human actions that our motion parts are directional.

Algorithm 1 Extract Representation from STIPs

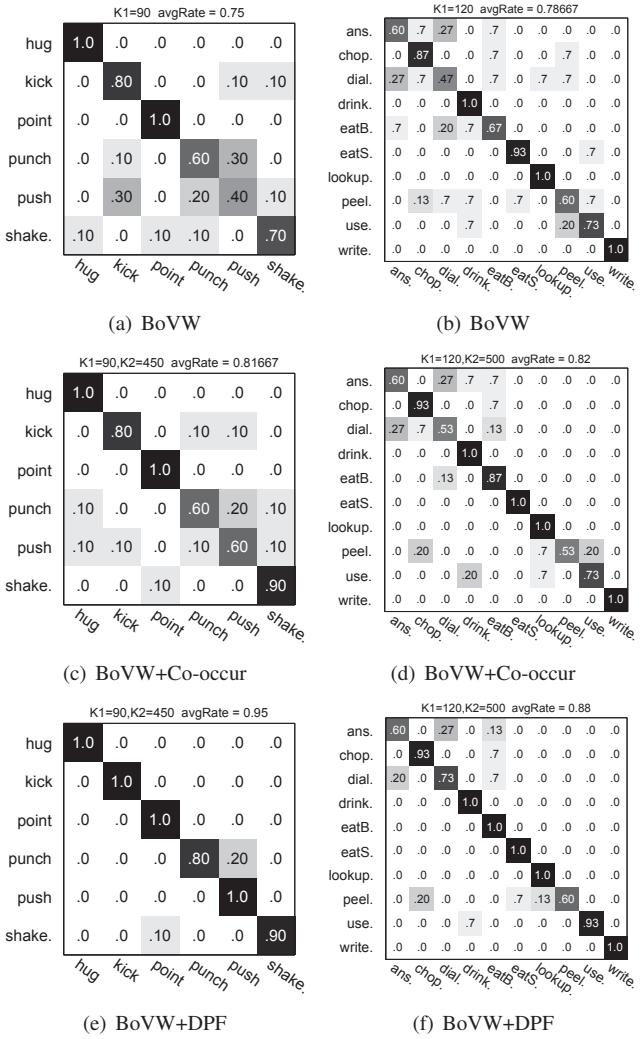
Require: Video $= \{I_t\}_{t=1}^{F}$, frame number *F*, constant *K*.
Ensure: histogram H

```
compute STIPs: S = {(x,y,t) | (x,y) ∈ I_t, 1 ≤ t ≤ F}
  and descriptors:{des_(x,y,t)}
cluster {des_(x,y,t)} into K centers, N(k) is the number
  of STIPs labeled k (k = 1,...,K) , S is divided into
  {S_1, S_2, ..., S_K}, T is calculated by Formula (2)
for i = 1 to K, j = 1 to K do
    if i ≠ j then
        for ∀pt_i ∈ S_i, pt_j ∈ S_j do
            calculate n(pt_i, pt_j) using Formula (1) and T
        end for
        count N(i,j) by Formula (3)
    end if
end for
for i = 1 to K do
    obtain P(DPF_i^st|DPFs) and P(DPF_i^en|DPFs)
    from Formula (4),(5)
end for
calculate H by Formula (6)
return H
```

3. EXPERIMENTS AND DISCUSSIONS

The proposed descriptor is evaluated on two challenging datasets: UT-Interaction [11] and Rochester [12]. Segmented version of UT-Interaction is utilized which contains 6 categories: "hug", "kick", "point", "punch", "push" and "shake-hands" [13]. "Point" is performed by single actor and other actions are performed by actors in pairs. All actions are repeated 10 times in two scenes resulting in 120 videos. Scene-1 is taken in a parking lot with little camera jitter and slightly zoom rates. In scene-2, the backgrounds are cluttered with moving trees, camera jitters and passerby. Rochester dataset contains 150 videos of 5 actors performing 10 actions: "answer a phone", "chop a banana", "dial a phone", "drink water", "eat a banana", "eat snacks", "look up a phone number in a phone book", "peel a banana", "eat food with silverware" and "write on a white board". Each action is repeated 3 times in the same scenario.

This work applies Laptev's detector [14] obeying the original parameter sets to detect STIPs and uses HOG [15] to generate 90 dimension descriptors. After extracting 800 points from each video, K-means clustering is applied to generate visual words, with 450 words for UT-Interaction (scene-1,scene-2) and 500 words for Rochester. Recognition is conducted using a non-linear SVM with a chi-squared kernel [16]. A leave-one-out cross validation is adopted for training-testing. Since random initialization is involved in



(a) BoVW (b) BoVW

(c) BoVW+Co-occur (d) BoVW+Co-occur

(e) BoVW+DPF (f) BoVW+DPF

Fig. 3. Confusion matrices for scene-1 of UT-Interaction in first column and for Rochester in second column. From top to down, BoVW, BoVW+Co-occurrence Map and BoVW+DPF are applied. *K1* is cluster number for BoVW and *K2* is used for Co-occurrence Map and DPF.

the K-means clustering method, all confusion matrices are average values over 10 times running results.

Experiments on UT-Interaction scene-1 and Rochester are shown in Fig.3. In each column of Fig.3, DPF and Co-occurrence Map combing with BoVW are compared with original BoVW using confusion matrices. In UT-Interaction scene-1, cluster number $K1$ is set 90 for BoVW. Most errors happens among "punch", "push" and "shake-hands" in (a). Co-occurrence Map in (c) slightly improves the discrimination of "push" and "shake-hands" by adding spatial-temporal information. DPF in (e) outperforms (a) and (c) and obtains the highest recognition rate. The reason lies in its abilities to capture directional spatial information of "punch", "push" and "kick" comparing with Co-occurrence Map. Considering vertical position between two points located on action executor's foot and action receiver's thigh, it changes for "kick"(choose the stretched out foot) while keeps unchanged for "push". DPF also reduced the errors among "answer phone", "dial phone" and "eat a banana" in Rochester since extra spatial information is encoded.

Table I. Compare proposed method with state-of-the arts.

UT-Interaction	scene-1/K	scene-2/K	Rochester/K
Dollar, *et al.*[1]	58.13%/90	45.06%/90	–
Sun, *et al.*[9]	82.67%/120	79.22%/120	–
Ryoo [5]	88%/800	77.00%/800	–
Liu, *et al.*[17]	85.00%/800		–
Satkin, *et al.*[18]	–	–	80.00%/4000
Messing, *et al.*[12]	–	–	89.00%/400
BoVW	75.00%/90	76.67%/90	78.67%/120
BoVW+Co-occur	81.67%/450	78.33%/450	82.00%/500
BoVW+DPF	**95.00**%/450	**86.67**%/450	**88.00**%/500

Table I compares the performance of proposed method with state-of-the-arts and cluster number K is marked with classification rate. Since parameters like the number K of K-means clustering method differs in different algorithms, the accuracy refers the classification rate with optimal parameters. The results on UT-Interaction are most directly comparable to the method in [1] and [9]. Our BoVW shows 16.87% and 31.61% higher than [1] which also obeys basic BoVW framework since Laptev's STIPs detector and descriptor are adopted. Our BoVW+DPF achieve average accuracies of 95.00% on UT-Interaction scene-1 and 86.67% on scene-2. Improvements of 12.33% and 7.45% are respectively achieved over [9], which can be attributed to our addition of directional spatial information. Since [5] mainly infers ongoing activities from partial videos, experiments are conducted on [5] with full observation, which still shows lower accuracies (7% lower for scene-1 and 9.67% for scene-2) than our method. Noticing backgrounds in the scene of Rochester are still, STIPs can be refined using background subtraction. This refinement is not included since our experiments focus on proving the ability of DPF comparing with traditionally pairwise features. The result is still competitive with [12] without using STIPs selection.

4. CONCLUSIONS

In this work, we propose a new descriptor called directional pairwise feature (DPF) to classify videos containing confusing human actions. Different with BoVW based methods and related works in capturing structural information, DPF involves the words' co-occurrence statistic as well as their directional information. Additionally, a dimension reduction method is applied to form the final action descriptor. Since richer information of spatial-temporal distribution is involved, DPF outperforms most BoVW based methods and the state-of-the-arts on two challenging datasets. Experiment results prove the robustness and efficiency of DPF against cluttered backgrounds and inter-class action ambiguities.

5. REFERENCES

[1] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *VS-PETS, pp.65-72, 2005.*

[2] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC, pp.124.1-124.11, 2009.*

[3] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *ACM Conf. Multimedia, pp.357-360, 2007.*

[4] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," in *BMVC, vol.3, pp.1249-1258, 2006.*

[5] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *ICCV, pp.1036-1043, 2011.*

[6] G. J. Burghouts and K. Schutte, "Spatio-temporal layout of human actions for improved bag-of-words action detection," in *Pattern Recognition Letters, 2013.*

[7] M. S. Ryoo and Aggarwal J. K., "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *ICCV, pp.1593-1600, 2009.*

[8] S. Savarese, A. DelPozo, J.C. Niebles, and L. Fei-Fei, "Spatial-temporal correlatons for unsupervised action classification," in *WMVC, pp.1-8, 2008.*

[9] Q. Sun and H. Liu, "Action disambiguation analysis using normalized google-like distance correlogram," in *ACCV, 2012, Part III, LNCS 7726, pp.425-437, 2013.*

[10] Q. Sun and H. Liu, "Learning spatio-temporal co-occurrence correlograms for efficient human action classification," in *ICIP, pp.3220-3224, 2013.*

[11] M. S. Ryoo, C. C. Chen, J. K. Aggarwal, and et al., "An overview of contest on semantic description of human activities (sdha) 2010," in *Recognizing Patterns in Signals, Speech, Images and Videos, pp.270-285, 2010.*

[12] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *ICCV, pp.104-111, 2009.*

[13] J. M. Carmona and E. J. Fernndez-Caballero, "A survey of video datasets for human action and activity recognition," in *CVIU, vol.117, Issue 6, pp.633-659, 2013.*

[14] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *ICPR, pp.32-36, 2004.*

[15] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Object detection with discriminatively trained part-based models," in *PAMI, vol.32, pp.1627-1645, 2010.*

[16] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *http://www.vlfeat.org/, 2008.*

[17] H. Liu, R. Feris, and M. T. Sun, "Benchmarking human activity recognition," in *CVPR Tutorial, CVPR, 2012.*

[18] S. Satkin and M. Hebert, "Modeling the temporal extent of actions," in *ECCV, pp.536-548, 2010.*