BCA: BI-SYMMETRIC COMPONENT ANALYSIS FOR TEMPORAL SYMMETRY IN HUMAN ACTIONS

Chenyang Zhang and Yingli Tian

The City College of New York,
{czhang10@citymail, ytian@ccny}.cuny.edu

ABSTRACT

In the past, many research efforts are invested into discriminative action recognition task but the general temporal structure of human actions is overlooked. In this paper, we focus on a specific yet common structure of human actions: temporal symmetry. The key contribution is that we model the temporal symmetry property of human action and separate this signal out of original action sequences without specifying which action category. Based on this modeling, a novel and effective method is proposed to detect the temporal symmetric part of any given human action sequence. Experimental results on two popular human action datasets verify that the temporal symmetry benefits both action detection and action recognition.

Index Terms— Bi-symmetric, Action Detection

1. INTRODUCTION

Video understanding is an essential application in multimedia research and application areas. Human activity recognition plays a very significant role in video understanding. To make human activities detectable by computer algorithms, researchers have proposed numerous models to describe human activities in many aspects, such as postures [1, 2], shapes [3, 4, 5], motions [6, 7, 8] and local appearances [9, 10].

However, most of the previous action recognition algorithms treat the problem as "video classification", where a compact video representation (such as a feature vector) is computed based on all the contents inside the video and then a classifier is trained and applied to map the representation vector to the class label. This framework is suitable for controlled settings (where the start and end frames of the actions are known) but problematic when the action of interest occupies only a small and unknown portion of the whole video. In this situation, action recognition needs the help from action detection, which provides a reasonable estimation of the spatial and temporal locations of the action of interest [12, 13, 14, 15, 16].

Previous algorithms handle the action detection (or localization) within the scope of action recognition, *i.e.*, firstly action-specific templates or classifiers are modeled as action templates; then the templates are used to probe that action in test video sequences by matching [17, 18]. To the best of our knowledge, there is no previous work trying to find an action independent detector only by exploring its temporal structure.

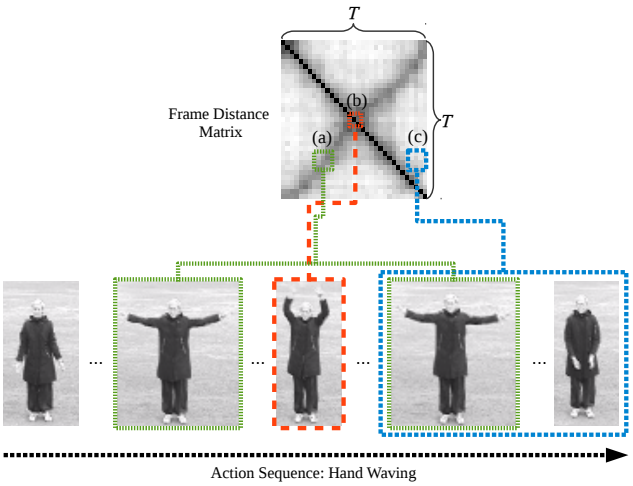


Fig. 1. An example of action temporal symmetry ("Waving") from the KTH dataset [11]. This symmetry pattern can be visualized and characterized by a frame-to-frame distance matrix in some feature space (Frame Distance Matrix). The temporal symmetric property can be characterized and visualized by the Bi-symmetric property of this distance matrix as well as the dark anti-diagonal in the distance matrix. (a) shows a pair of frames which are temporal symmetric around (b) the pivot frame, and their corresponding entries locate on the dark anti-diagonal, and (c) shows a pair of frames which are distant from each other.

Detecting human actions with pre-defined action templates has three drawbacks: 1) it needs a large dataset to train a reliable action-specific detector because different subjects perform the same action differently and even the same subject performs differently under different scenarios. 2) For real-time action detection systems, such as surveillance systems, the complexity to detect an action is proportional to the number of pre-defined action classes, which is inefficient when the number of action classes is large. 3) The whole system needs to be retrained if new action classes are added. Therefore, it is significant to explore possible solutions to detect temporal

extents of actions of interest without pre-defined action templates and instead using the temporal structure directly.

This paper focuses on a specific yet very common intrinsic property of human actions: temporal symmetry. We concede that although not every action class contains temporal symmetric, there are many common action classes are (at least partially) temporal symmetric. As shown in Fig 1, action "Waving" shows strong temporal symmetry pattern and the temporal pivot of symmetry is the frame where hands are raised to the apex position.

Our key observation is that if an action sequence is temporal symmetric, its frame-to-frame distance matrix has the following characteristics: 1) it is bi-symmetric, *i.e.,* symmetric around both diagonal and anti-diagonal. 2) Both diagonal and anti-diagonal have low-intensity entries. The frame-to-frame distance matrix in Fig 1 visually characterize these patterns. The contributions of this paper are two-folded:

- A new temporal symmetry structure of human actions and apply it to human detection and recognition tasks.
- A mathematical model of the temporal symmetry pattern is proposed by leveraging frame-distance matrix and a novel quantitative method is proposed to video analysis tasks.

Section 2 reviews some existing work about action recognition and detection. Section 3 describes the proposed model of the temporal symmetry structure and the quantitative method for applying to video analysis tasks. Then the evaluation of the proposed model on several public datasets in action recognition and detection is presented in Section 4. Section 5 concludes the paper.

2. RELATED WORK

Human action detection and recognition play important roles in many computer vision applications such as video surveillance [15, 14] and media retrieval [12]. In [12], combination of shape and motion cues is investigated in a "drinking" action recognizer to detect its spatial-temporal extents in movie segments. In surveillance videos, since the backgrounds are more complex and cluttered, human detection and tracking are employed to generate hypothesis for further processes [15]. Besides the different feature extraction and hypothesis proposal methods, these models share similar structures with other action recognition frameworks. Different from the detect-and-recognize framework, Yuan *et al.* proposed to use a recognize-and-detect framework: 1) Firstly each space-time interest point is assigned a class-dependent weight from a learning model. 2) Then an efficient 3D branch-and-bound algorithm is applied to search for an optimal 3D-bounding volume for that action. However, although these two kinds of methods are very different in their architectures, they did not explore the intrinsic temporal structures of human actions and are still dependent on action-specified contents. Different from previous methods, we focus on exploring a specific temporal structure of actions and propose an action independent detector.

3. PROPOSED METHOD

3.1. Problem Statement

We firstly review the definition of a persymmetric matrix. Following the convention in [19], a superscript F is used to refer to flip-transpose (transpose over the anti-diagonal). If a matrix M is equal to its flip transposed one, M^F, then M is a *Persymmetric* matrix. Similarly, if a matrix is both symmetric and persymmetric, it is *"Bisymmetric"*. Therefore, for a bisymmetric matrix M, there is:

$$M = M^T = M^F. \tag{1}$$

Given a video sequence V which is composed of T frames $V = \{I_1, I_2, ..., I_T\}$, where I_i represents the i^{th} image frame in the video. A video description generator (feature descriptor) $\phi(\cdot)$ is applied to each frame: $f_i = \phi(I_i)$.

In our work, the Bag-of-Words [20] model together with some local feature extractors are employed. The local feature extractors are different for different image modalities, more implementation details are described in section 4. After the feature extraction step, the description of the input video V can be described as:

$$\phi(V) = \{f_1, f_2, ..., f_T\}, \tag{2}$$

where all f_i is the descriptor vector of image frame i and all descriptors across all frames are of the same dimension.

Therefore, if the video sequence is exactly temporal symmetric, there is:

$$\|f_i - f_{T-i+1}\| = 0, \forall i = 1, 2, ...T. \tag{3}$$

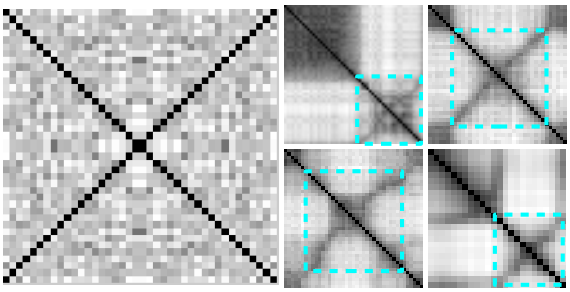
Subsequently, a frame-to-frame distance matrix D can be computed by:

$$D_{ij} = \|f_i - f_j\|, \forall i, j = 1, 2, ...T. \tag{4}$$

Thus, based on the properties of *Bisymmetric* matrix, an ideal temporal symmetric sequence's frame-to-frame distance matrix is perfectly *Bisymmetric*, and the elements on both of its diagonal and anti-diagonal are zeros.

As illustrated in Fig 2, if a video sequence is ideally temporal symmetric, the frame distance matrix D should be like the one in (a), where only the two diagonals are zeros and elsewhere are non-zeros, and it is bisymmetric.

In reality, due to variance in appearances and different temporal structures (such as offset and different execute rates) of the action performed by a real human being, D may be different from the perfect one in different videos. For example, some different distance matrices are visualized in Fig 2 (b).



(a) Ideal (b) Reality

Fig. 2. Illustration of frame distance matrices of (a) an ideal temporal symmetric sequence and (b) real feature sequences from samples from MSR Action3D Dataset [21]. Cyan dashed boxes show the sub-matrices which most follow the bisymmetric properties.

We observe some facts: 1) the anti-diagonal is not all-zeros, but it is much darker than other entries. 2) The dark "anti-diagonal" may not align perfectly with the real anti-diagonal due to variances in action execution rates and appearances. 3) Instead of the whole distance matrix, a submatrix of it may be more suitable to be considered as a bisymmetric matrix due to different starting and ending time of the action execution.

Although the frame distance matrix is not perfect, we can still observe the two visual patterns from them: 1) near bisymmetric and 2) a dark anti-diagonal in parts labeled by cyan dashed boxes. In this paper, we propose a method called "Bisymmetric Component Analysis"(BCA) to find the most temporal symmetric segment by detecting the most bisymmetric sub-matrix.

3.2. Bisymmetric Component Analysis

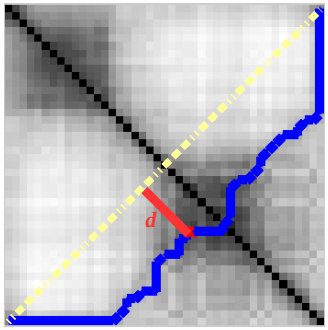


Fig. 3. Illustration of "off-diagonal" cost. Yellow dashed line is the anti-diagonal and blue line is the weighted-shortest path from bottom-left to top-right. For each point on that path, the off diagonal distance is d (colored in red).

Sliding windows of varied sizes are employed to generate a possible temporal proposal set:

$$P = \{p_i\}, \;\; p_i = (s_i, e_i, D_i), \tag{5}$$

where the tuple (s_i, e_i, D_i) indicating the starting/ending frames and corresponding sub-matrix of proposal p_i. To find the top K proposals among all, we first rank them by a scoring function and then discard a portion of them using an eigenvalue property of bisymmetric matrices.

3.2.1. *Ranking the Temporal Proposals*

The first scoring term is defined based the fact that the sub-matrix should be persymmetric:

$$s_1(p_i) = \|D_i - D_i^F\|_F^2 \tag{6}$$

The second scoring term is based on the fact that the elements of the sub-matrix near the off-diagonal should be close to zeros, as visualized in Fig 2 (b) and Fig 3. The weighted-shortest path is first computed from bottom-left to top-right of each D_i, the cost of such a path is the summation of matrix values the path covers. Because a "dark" anti-diagonal is preferred for a good D_i, the following scoring term is employed to model this:

$$s_2(p_i) = (\Sigma_{j \in \epsilon_i} d_j)/|\epsilon_i|, \tag{7}$$

while ϵ_i is the shortest path of D_i and d_j is the off-diagonal distance for each pixel j on ϵ_i, as illustrated in Fig 3. To rank all proposals, a weighed sum of s_1 and s_2 is employed: $s(p_i) = w_1 s_1(p_i) + (1 - w_1) s_2(p_i)$. In our experiments, we set w_1 to 0.5.

3.2.2. *Temporal Proposal Purification*

The proposed submatrices are further purified by selecting those whose properties are more compatible with the bisymmetric matrices' properties elaborated as follows.

As proved in [19], for bisymmetric matrices, we have the following theorem:

THEOREM 1 *The eigenvalues of a bisymmetric matrix G, which have the following structure:*

$$G = \begin{bmatrix} A^F & B^T \\ B & A \end{bmatrix} \tag{8}$$

where A is symmetric and B is persymmetric matrices such that the eigenvalues of G are also the eigenvalues of $A + BR$ and $A - BR$. R has ones along the anti-diagonal and zeros elsewhere. Whenever v is an eigenvector of $A + BR$, $[v^T R^T, v^T]^T$ is an eigenvector of G with the same eigenvalue.

For more detailed proof of this theorem, please see [19].

Based on this theorem, we can evaluate submatrices proposed from the last step by leveraging the number of matched eigenvalue eigenvector pairs as a scoring function to evaluate how a matrix G follows the previous stated theorem:

$$\begin{aligned} &\forall (v, \lambda) \in eig(G), (v', \lambda') \in eig(A + BR): \\ &c((v, \lambda), (v', \lambda')) = g([v'^T R^T, v'^T]v) \times g(1 - \frac{2|\lambda - \lambda'|}{|\lambda| + |\lambda'|}), \end{aligned} \tag{9}$$

where $c(\cdot)$ is the cost function between pairs.

A small threshold is set for $c(\cdot)$ for matching and all proposals where there are less than 50% matched Eigen pairs are discarded. In this way, a "good" submatrix proposal has a high score in this function is ensured because most of its Eigen value and Eigen vector pairs can be found in corresponding matrix $A + BR$.

4. EXPERIMENTAL RESULTS

4.1. Datasets

MSR Action Dataset: MSR Action dataset [13] is composed of 16 video sequences and 63 action segments: 14 hand clapping, 24 hand waving, and 25 boxing, performed by 10 subjects. Both Indoor and outdoor scenes are addressed and some are captured under clutter moving backgrounds. This dataset is selected to evaluate action detection result for two reasons: 1) the action classes (*e.g.,* hand waving) have temporal symmetric properties and 2) studying temporal symmetric property in clutter backgrounds is more realistic.

MSR Action3D Dataset: MSR Action3D Dataset [21] is a depth-based dataset captured by a Kinect camera. There are 20 gaming-related action categories ranging from "Two-arm Waving" to "Golf-swing". There are also 10 subjects involved in this dataset and each subject performs each action 2 or 3 times. There are 567 depth video sequences in total. The resolution is also 320×240. This dataset contains relative more action classes than the MSR Action Dataset. This dataset is employed to evaluate how BCA will benefit action recognition. Although this dataset is not as realistic as **MSR Daily** [22], MSR Action3D is designed for gaming applications and contains more temporal symmetric actions, which is more suitable to our study than **MSR Daily**. Sample frames from the datasets and the temporal symmetric property of actions are illustrated in Fig 4.

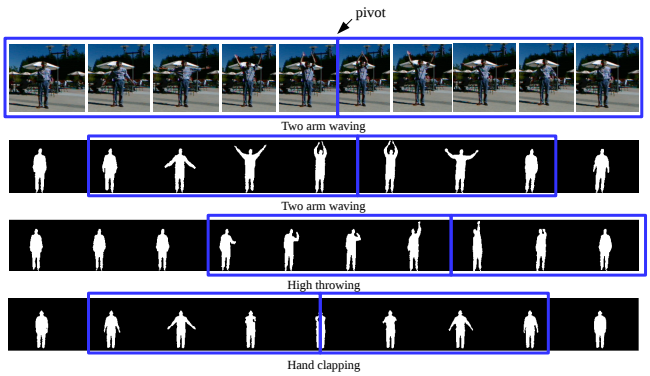


Fig. 4. Sampled frames from MSR Action Dataset(the top row) and MSR Action3D Dataset (the rest). Temporal extents of symmetry and pivot frame are also illustrated.

4.2. Action Detection

Firstly, the proposed BCA method is evaluated in action detection using the MSR Action Dataset. The labeled spatial bounding boxes are used to generate 63 test video sequences which focus on the subject performing the action of interest. The reason of utilizing the bounding boxes is that instead of detecting actions in spatial extents, this paper focuses more on detecting in temporal extents.

STIP [10] is densely extracted from the video sequences and HOG-HOF descriptor is used for video description. The K-means algorithm is employed to generate a visual vocabulary of 3000 visual words. Therefore, each frame is represented by a bag-of-words (BoW) vector of 3000 elements. The frame-to-frame distance matrix of each video sequence is calculated by the l_2 distance between BoW vectors of each pair of frames.

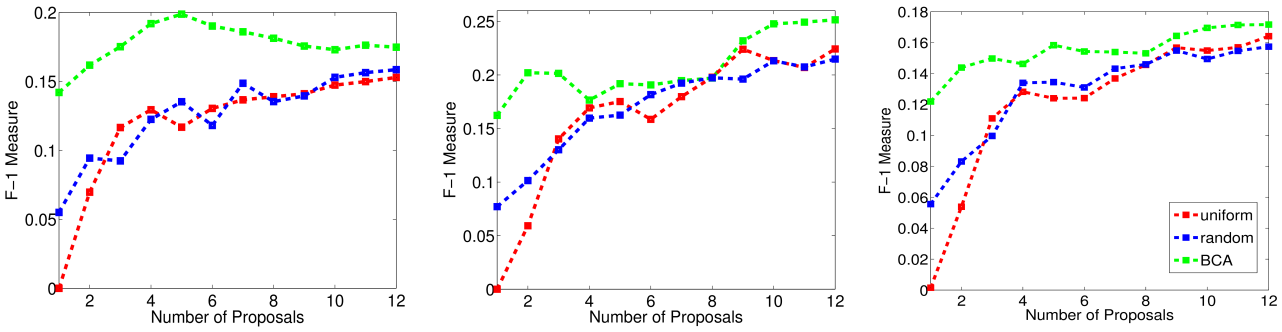
After applying description on video sequences, the proposed bisymmetric component analysis method (BCA) is employed to find a subsequence of the video sequence which complies mostly with the bisymmetric property, which will be treated as the temporal extent detection of an action.

To quantitatively evaluate the detection results, the F-1 measures are computed based on detection precision and recall. The top K temporal proposals are selected and the union frames of them is used to compare with the ground-truth. Since to our best knowledge there is no general detection methods reported before, we compare our results with two baselines: randomly and uniformly sampling temporal proposals. For random sampling, the evaluation is repeated 10 times and the average numbers of 10 runs are reported. The results are visualized in Fig 5. When set $K = 1$, which means only the top proposal is accepted, our method is advantageous to the baselines by a large margin. With the increased number of proposals, two baselines almost meet but our method (BCA) is consistently advantageous.

4.3. Action Recognition based on Detection

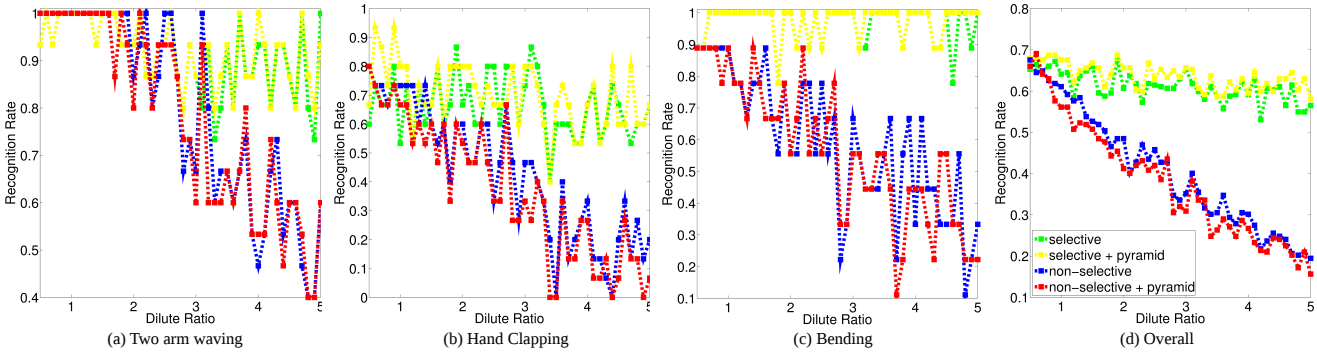
To evaluate whether action detection can help action recognition, we conduct experiments on the MSR Action 3D Dataset. H3DF features [23] are densely sampled and BoW representation are generated in the same scheme as in the previous experiment. Since the dataset is not collected for detection, we create 46 "diluted" versions of this dataset by adding random frames from the same subject but different actions before and after the original sequences. The dilute ratio ranges from 0.5 to 5, which means we randomly add 50% to 500% random frames to the original videos.

Under such a setting, the detection becomes very critical to the final action recognition results because with up to 500% noise, the true signal is easily overwhelmed. The recognition is based on max-pooling and linear SVMs. Compared with baseline "non-selective" as in Fig 6, it is easy to find that if there is no such detection, the recognition rates drop very fast



(a) Boxing (b) Waving (c) Overall

Fig. 5. Action detection scores in F-1 measures on the MSR Action Detection Dataset. We compare our method (BCA) with two baselines using uniform sampling and random sampling methods, respectively. (a) and (b) are for actions "Waving" and "Boxing" only while (c) shows the average scores over all three action categories (action "Clapping" is not shown here since the movement of this action is very subtle to observe the symmetry pattern.) The margin of performances of BCA over baselines is the largest when only the top video segment proposal is used. With the number of proposals increases, the margin decreases to some degree but BCA is advantageous consistently.



(a) Two arm waving (b) Hand Clapping (c) Bending (d) Overall

Fig. 6. Action recognition rates over increasing dilute ratios. Green and yellow curves show results with the help of the proposed detection results with and without temporal pyramids, respectively. Blue and red curves show results without detection. (a),(b) and (c) show action classes "two-arm waving", "hand clapping" and "bending", respectively. (d) shows the overall results. With the increasing dilute ratios (the detection gets harder), the proposed detection generates stable results while recognition without detection decreases quickly.

with increasing dilute ratio, however, our method performed consistently well (Fig 6 (d)) and there is only small decrease in recognition rates.

To further investigate the discovered temporal structures, we also combine the proposed method and baseline method with the well-known temporal pyramid to capture some structure information, in this experiment, we apply two layers pyramids. The performances are illustrated as yellow and red curves in Fig 6 (d). We can observe that if the temporal location is accurately detected, adding temporal pyramid helps adding more information (yellow curve is above green curve). If there is no accurate detection, adding pyramids can only bring more noise (red curve is under blue curve). This observation further demonstrates the accuracy of the proposed detection algorithm.

5. CONCLUSION

This paper have addressed the following question: can we detect an action without knowing which class it is? While traditional action detection can only detect a specific type of action by probing the segments from video sequences using a pre-trained template or classifier, we have proposed a general action detector focusing on the temporal symmetry pattern without specifying action category. We concede that not all action classes are suitable for symmetry detection, but at least the temporal symmetry is a very common pattern in many actions. With experiments on two popular action datasets, we have demonstrated the validity of the proposed BCA algorithm based on frame-to-frame distance matrices and observed that the detection contributes to action recognition. Our future work will be further discovering other temporal patterns (*e.g.* repetitive.)

Acknowledgement

This work was supported in part by NSF grants EFRI-1137172, IIP-1343402, and IIS-140080.

6. REFERENCES

[1] Junji Yamato, Jun Ohya, and Kenichiro Ishii, "Recognizing human action in time-sequential images using hidden markov model," in *CVPR*. IEEE, 1992, pp. 379–385.

[2] Liang Wang and David Suter, "Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model," in *CVPR*. IEEE, 2007, pp. 1–8.

[3] Aaron F. Bobick and James W. Davis, "The recognition of human movement using temporal templates," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 3, pp. 257–267, 2001.

[4] Daniel Weinland, Remi Ronfard, and Edmond Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 249–257, 2006.

[5] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri, "Actions as space-time shapes," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*. IEEE, 2005, vol. 2, pp. 1395–1402.

[6] Jens Rittscher and Andrew Blake, "Classification of human body motion," in *ICCV*. IEEE, 1999, vol. 1, pp. 634–639.

[7] Ramprasad Polana and Randal Nelson, "Low level recognition of human motion (or how to get your man without finding his body parts)," in *Motion of Non-Rigid and Articulated Objects, 1994., Proceedings of the 1994 IEEE Workshop on*. IEEE, 1994, pp. 77–82.

[8] Alexei A Efros, Alexander C Berg, Greg Mori, and Jitendra Malik, "Recognizing action at a distance," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 726–733.

[9] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie, "Behavior recognition via sparse spatio-temporal features," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*. IEEE, 2005, pp. 65–72.

[10] Ivan Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[11] Ivan Laptev and Tony Lindeberg, "Velocity adaptation of space-time interest points," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. IEEE, 2004, vol. 1, pp. 52–56.

[12] Ivan Laptev and Patrick Pérez, "Retrieving actions in movies," in *ICCV*. IEEE, 2007, pp. 1–8.

[13] Junsong Yuan, Zicheng Liu, and Ying Wu, "Discriminative subvolume search for efficient action detection," in *CVPR*. IEEE, 2009, pp. 2442–2449.

[14] Ming Yang, Fengjun Lv, Wei Xu, Kai Yu, and Yihong Gong, "Human action detection by boosting efficient motion features," in *ICCV Workshops*. IEEE, 2009, pp. 522–529.

[15] Ming Yang, Shuiwang Ji, Wei Xu, Jinjun Wang, Fengjun Lv, Kai Yu, Yihong Gong, Mert Dikmen, Dennis J Lin, and Thomas S Huang, "Detecting human actions in surveillance videos," in *TREC Video Retrieval Evaluation Workshop*, 2009.

[16] Qingshan Luo, Xiaodong Kong, Guihua Zeng, and Jianping Fan, "Human action detection via boosted local motion histograms," *Machine Vision and Applications*, vol. 21, no. 3, pp. 377–389, 2010.

[17] Dong Huang, Shitong Yao, Yi Wang, and Fernando De La Torre, "Sequential max-margin event detectors," in *ECCV 2014*, pp. 410–424. Springer, 2014.

[18] Victoria Bloom, Dimitrios Makris, and Vasileios Argyriou, "G3d: A gaming action dataset and real time action recognition evaluation framework," in *CVPR Workshops*. IEEE, 2012, pp. 7–12.

[19] Gene H Golub and Charles F Van Loan, *Matrix computations*, vol. 3, JHU Press, 2012.

[20] Josef Sivic and Andrew Zisserman, "Efficient visual search of videos cast as text retrieval," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 4, pp. 591–606, 2009.

[21] Wanqing Li, Zhengyou Zhang, and Zicheng Liu, "Action recognition based on a bag of 3d points," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 9–14.

[22] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *CVPR*. IEEE, 2012, pp. 1290–1297.

[23] Chenyang Zhang and Yingli Tian, "Histogram of 3d facets: A depth descriptor for human action and hand gesture recognition," *Computer Vision and Image Understanding*, 2015.