# MULTI-CHANNEL CORRELATION FILTERS FOR HUMAN ACTION RECOGNITION

*Hamed Kiani*[⋆] *Terence Sim*[⋆] *Simon Lucey*[†]

[⋆] School of Computing, National University of Singapore [†] Carnegie Mellon University
[⋆] {hkiani, tsim}@comp.nus.edu.sg [†] slucey@andrew.cmu.edu

## ABSTRACT

In this work, we propose to employ multi-channel correlation filters for recognizing human actions (e.g. *waking*, *riding*) in videos. In our framework, each action sequence is represented as a multi-channel signal (frames) and the goal is to learn a multi-channel filter for each action class that produces a set of desired outputs when correlated with training examples. The experiments on the Weizmann and UCF sport datasets demonstrate superior computational cost (real-time), memory efficiency and very competitive performance of our approach compared to the state of the arts.

***Index Terms***— Action recognition, Correlation filters, Multi-channel features

## 1. INTRODUCTION

Human action recognition is a challenging problems in computer vision which has received substantial attention over the last few years. In general, the difference of current approaches mainly comes form the basis of the representation used for actions. Some leading representations are learned geometrical models of human body parts [1], space-time pattern templates, appearance or region features, shape or form features [2] [3], interest-point-based representations [4], volumetric features [5], and motion/optical flow patterns [6].

Correlation filters, developed initially in the seminal work of Hester and Casasent [7], are a method for learning a filter in the frequency domain that returns corresponding desired outputs when correlated with a set of training signals (e.g. images) [7, 8, 9]. Interest in correlation filters has been reignited in the vision world through the recent work of Bolme et al. on Minimum Output Sum of Squared Error (MOSSE) correlation filters [10]. This work addressed some of the classical problems with earlier correlation filters (e.g. over-training and poor generalization) and was extremely efficient in terms of computation and memory usage. Despite the great progress, traditional correlation filters have been rarely applied on challenging pattern detection/recognition with large inter-class similarities and intra-class variations, due to their inability to handle modern descriptors (e.g. HOG [11]) for discriminative filter training. More recently, Kiani et al. [12] introduced Multi-Channel Correlation Filters (MCCFs) to employ discriminative descriptors such as HOG for learning multi-channel filters/detectors efficiently in the frequency domain. They demonstrated competitive results of MCCFs across a myriad of challenging detection/recognition tasks with superior computational and memory efficiency in comparison to spatial detectors.

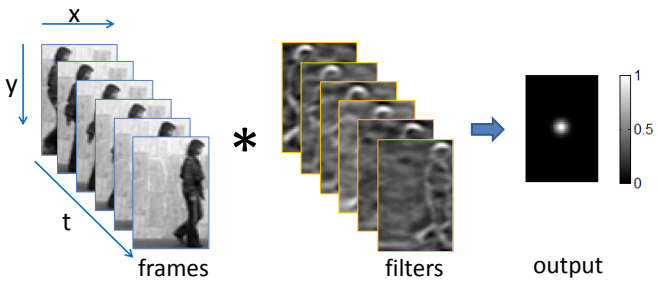


**Fig. 1**. An example of learning a multi-channel correlation filter for *walking* action. The action cycle is represented by $N$ intensity frames and the goal is to learn an $N$-channel correlation filter that returns a desired correlation output (a 2D Gaussian) when correlated with the training action cycle.

The application of correlation filters has recently been investigated for human action detection/recognition with promising results [13] [3] [14]. The main idea of these approaches is to represent actions using spatio-temporal volumes and learn a volume correlation filter in the 3D frequency domain that produces a peak at the origin of the action in both spatial and temporal domain. More specifically, Rodriguez et al. [3] has extended the Optimal Trade-off Maximum Average Correlation Height (OT-MACH) filter to 3D MACH and proposed action MACH to train 3D correlation filters for action recognition in video sequences. The main advantage of this approach is its closed-form solution for both scaler and vector features which makes training process computationally very efficient. Moreover, detection can be made extremely fast due to the efficiency of correlation in the frequency domain.

This method, however, suffers from some major limitations. First, action MACH trains a correlation filter that satisfies a set of criteria (e.g. maximizing the average correlation height) over all positive training examples. It has been shown by Ali and Lucey [13] that action filters trained using action

MACH are equivalent to the average of the action specific examples which may suffer from poor generalization for unseen data and over-training for training examples. Second, action MACH only makes use of positive examples and ignores negative examples during learning process (according to its leaning objective). This may result in training correlation filters with low discrimination power which perform poor against large inter-class similarities (confusions among *walking*, *jogging* and *running* in [3]). Finally, action MACH does not specify desired values over the entire correlation outputs of training examples, as all supervised learning techniques basically do. It was discussed in [10] that this may increase the sensitivity to the noise or produce smooth peaks which are difficult to be accurately recognized/detected.

In this paper, we proposed to employ multi-channel correlation filters [12] for human action recognition in videos. The core idea is that each action example with $N$ time-ordered frames can be considered as a multi-channel signal (with $N$ channels for scalar features such as image intensity and $N \times M$ channels for vector features like $M$ bins HoG). Given a set of training examples and their corresponding correlation outputs, the goal is to learn a multi-channel action filter in the frequency domain that produces the desired correlation outputs when correlated with the training examples (Figure 1).

**The advantages of MCCFs** for action recognition are as follows. First, the ridge regression form of MCCFs objective in the spatial domain [12] allows us to specify the desired values for the entire correlation outputs. This significantly reduces the instability against the noise and practically produces sharp peaks for more accurate detection/recognition. Second, the MCCFs is capable of exploiting both positive and negative examples for discriminative filter training. Finally, filter training and testing can be performed very efficiently in the frequency domain for both scalar and vector data.

## 2. CORRELATION FILTERS

The MOSSE filter [10] can be expressed as solving the following ridge regression problem in the spatial domain,

$$E(\mathbf{h}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{D} ||\mathbf{y}_i(j) - \mathbf{h}^\top \mathbf{x}_i[\Delta\boldsymbol{\tau}_j]||_2^2 + \frac{\lambda}{2}||\mathbf{h}||_2^2 \quad (1)$$

where $\mathbf{y}_i \in \mathbb{R}^D$ indicates the desired response for the $i$-th example $\mathbf{x}_i \in \mathbb{R}^D$ and $\lambda$ is a regularization term. $\mathbb{C} = \{\Delta\boldsymbol{\tau}_j\}_{j=1}^{D}$ represents the set of all possible circular shifts for a signal of length $D$. Solving Equation 1 in the spatial domain quickly becomes intractable respect to the signal length $D$, as it needs to solve a $D \times D$ linear system with a cost of $\mathcal{O}(D^3 + ND^2)$ [12]. It is well understood in signal processing that circular convolution in the spatial domain can be expressed as a Hadamard product in the frequency domain. Thus, Equation 1 can be equivalently expressed as,

$$E(\hat{\mathbf{h}}) = \frac{1}{2}\sum_{i=1}^{N} ||\hat{\mathbf{y}}_i - \hat{\mathbf{x}}_i \circ \text{conj}(\hat{\mathbf{h}})||_2^2 + \frac{\lambda}{2}||\hat{\mathbf{h}}||_2^2 \quad (2)$$

where $\hat{\mathbf{h}}, \hat{\mathbf{x}}, \hat{\mathbf{y}}$ are the Fourier transforms of $\mathbf{h}, \mathbf{x}, \mathbf{y}$. The complex conjugate of $\hat{\mathbf{h}}$ is used to ensure the operation is correlation not convolution. A solution to $\hat{\mathbf{h}}$ can be found with a cost of $\mathcal{O}(ND \log D)$ [12]. The primary cost is associated with the DFT on the ensemble of training signals $\{\mathbf{x}_i\}_{i=1}^{N}$ and desired responses $\{\mathbf{y}_i\}_{i=1}^{N}$.

## 3. MULTI-CHANNEL CORRELATION FILTERS

The objective of MCCFs in the spatial domain is defined as [12],

$$\begin{aligned} E(\mathbf{h}) &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{D} ||\mathbf{y}_i(j) - \sum_{k=1}^{K} \mathbf{h}^{(k)T}\mathbf{x}_i^{(k)}[\Delta\boldsymbol{\tau}_j]||_2^2 + \\ &\quad \frac{\lambda}{2}\sum_{k=1}^{K} ||\mathbf{h}^{(k)}||_2^2 \end{aligned} \quad (3)$$

where $\mathbf{x}^{(k)}$ and $\mathbf{h}^{(k)}$ refers to the $k$th channel of the vectorized image/frame and filter respectively where $K$ represents the number of filters. Solving this multi-channel form in the spatial domain is even more intractable than the single channel form with a cost of $\mathcal{O}(D^3K^3 + ND^2K^2)$ since one has to solve a $KD \times KD$ linear system.

Inspired by the efficiencies of posing single channel correlation filters in the frequency domain, Equation 3 can be expressed equivalently and more succinctly as,

$$\begin{aligned} E(\hat{\mathbf{h}}) &= \frac{1}{2}\sum_{i=1}^{N} ||\hat{\mathbf{y}}_i - \sum_{k=1}^{K} \text{diag}(\hat{\mathbf{x}}_i^{(k)})^T\hat{\mathbf{h}}^{(k)}||_2^2 + \\ &\quad \frac{\lambda}{2}\sum_{k=1}^{K} ||\hat{\mathbf{h}}^{(k)}||_2^2 \end{aligned} \quad (4)$$

where $\hat{\mathbf{h}} = [\hat{\mathbf{h}}^{(1)T}, \ldots, \hat{\mathbf{h}}^{(K)T}]^T$ is a $KD$ dimensional supervector of the Fourier transforms of each channel. This can be simplified further,

$$E(\hat{\mathbf{h}}) = \frac{1}{2}\sum_{i=1}^{N} ||\hat{\mathbf{y}}_i - \hat{\mathbf{X}}_i\hat{\mathbf{h}}||_2^2 + \frac{\lambda}{2}||\hat{\mathbf{h}}||_2^2 \quad (5)$$

where $\hat{\mathbf{X}}_i = [\text{diag}(\hat{\mathbf{x}}_i^{(1)})^T, \ldots, \text{diag}(\hat{\mathbf{x}}_i^{(K)})^T]$ and the solution for Equation 5 becomes,

$$\hat{\mathbf{h}}^* = (\lambda\mathbf{I} + \sum_{i=1}^{N}\hat{\mathbf{X}}_i^T\hat{\mathbf{X}}_i)^{-1}\sum_{i=1}^{N}\hat{\mathbf{X}}_i^T\hat{\mathbf{y}}_i \quad (6)$$

The cost of solving this linear system looks no different to the spatial domain as one still has to solve a $KD \times KD$ linear system. Fortunately, $\hat{\mathbf{X}}$ is sparse banded and it is shown in [12]

that Equation 5 can be efficiently solved through a variable re-ordering with smaller cost of $\mathcal{O}(DK^3 + NDK^2)$.

## 4. EXPERIMENTS

**Dataset:** We used two publicly available action datasets for evaluation: Weizmann dataset [2], and UCF sport dataset [3]. The Weizmann dataset contains 10 actions (bending, jumping, etc.) performed by 9 different subjects over a static background with slight changes of view point, scale and illumination. The UCF sport dataset is more challenging and contains 10 human actions such as diving and golf swinging filmed under challenging situations with background clutter, lightning/scaling changes, and significant intra-class variations.

**Features:** We evaluated our method using different features: normalized intensity, edge magnitude, temporal derivative and HOG (5 orientation bins normalized by cell and block sizes of $5 \times 5$ and $3 \times 3$, respectively). To compensate for the large illumination variation, all frames were power-normalized to have zero-mean and standard variation of 1.

**Desired Correlation Outputs:** For positive examples, a 2D Gaussian with spatial variance of 2 was employed to define the desired correlation outputs whose the peak was centered at the center of the last frame. A 2D matrix of zero values formed the desired correlation outputs of negative examples.

**Filter Training and Testing:** The annotations from [15] were used to extract training action cycles of both datasets. The cycles of each action class were carefully aligned in both spatial and temporal domains. Given a set of positive and negative training examples and their corresponding desired correlation outputs, the action specific filter was trained using Equation 6. For testing, we applied the MCCF filter trained for each class on a test video, and the label of the filter with maximum Peak-to-Sidelobe Ratio (PSR) [16] is assigned. To deal with scaling in the UCF dataset, a simple pyramid approach was employed to scan testing videos across different scales (from 0.4 to 1.5 of scaling-step 1.5) and the correlation output with maximum PSR across the pyramid was selected for each video. For actions with whole-body translation (e.g. *walking*) we trained two filters (left-to-right and right-to-left) by vertically flipping the training examples. We performed leave-one-subject-out cross-validation for the Weizmann dataset [2] and leave-one-sample-out for the UCF sport dataset.

**Quantitative Results:** The confusion matrix of our approach on the Weizmann dataset is illustrated in Figure 2 (top), showing 100% accuracy of our method for all action classes except *jump* and *skip*. Two confusions occurred between *jump* versus *skip* and *skip* versus *run* actions caused by their significant motion/appearance similarities. Figure 2 (bottom) shows our confusion matrix on the UCF sport dataset. The proposed method achieved promising results for most of the actions. There are more errors in *skating*, *running* and *kicking*. This

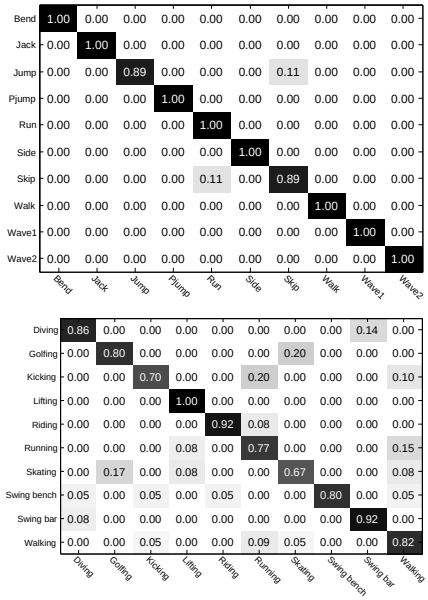


**Fig. 2**. The confusion matrix of our method for (top) the Weizmann, and (bottom) the UCF dataset(best viewed in pdf).

might be caused by the disadvantage of HOG to capture motion features in temporal domain which has been shown to be more robust against large motion similarities [17]. We can increase the robustness of MCCF against large inter-class motion similarities using more discriminative spatio-temporal features such as HOG3D [18] and HOG/HOF [4].

Table 1 provides a comparison of our method with those previously reported in the literature on the Weizmann and UCF sport datasets. For the Weizmann dataset, the highest mean recognition rate (100%) achieved by Huang et al. [19]. This method , however, was evaluated on 9 actions (*skip* was discarded) using a rich combination of optical flow and color histogram features. In addition, feature extraction, tracking and stabilization made Huang's method very slow. Our accuracy (97.8%) is slightly lower than this method on more action classes with real-time recognition speed, and higher than those reported by [20] and [17]. For the UCF dataset, our method achieved competitive accuracy compared to the stat-of-the-art. The best performance obtained by Cai et al. [21] using dynamic structure preserving map (DSPM) technique. It, however, suffers from heavy computation and sensitivity to video data redundancy. For both datasets, the action MACH [3] obtained the lowest performance of 86.6% (Weizmann) and 69.2% (UCF sport) due to sensitivity to inter-class similarities and poor generalization. Table 2 shows the robustness of our method against different types of features. Using these low level features can significantly reduce the feature dimension and processing time and, consequently, make recognition even more faster.

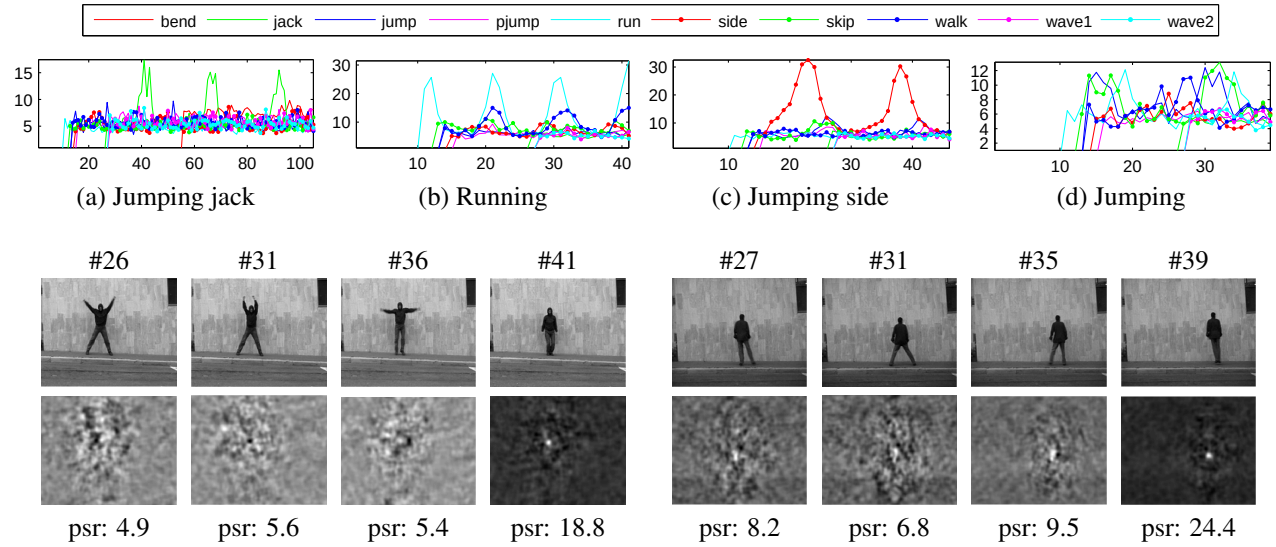


**Fig. 3**. Qualitative results. (top) PSRs versus frame number for some selected testing videos (best viewed in color). (bottom) Selected frames of two actions with corresponding correlation outputs. Please refer to **"Qualitative Results"** for explanation.

**Qualitative Results:** Figure 3 (top) shows the PSRs obtained by applying the trained action filters on some testing examples including *Jumping jack*, *Running*, *Jumping side* and *Jumping* (from the Weizmann dataset) versus frame numbers. Clearly, the PSRs produced by same-class filter are significantly higher then those produced by different-class filters. The *Jumping* is a failed case, where the PSRs produced by *Skip* filter over the test video is slightly higher than those by *Jump* filter. Interestingly, our method is able to produce high PSR for each action cycle through the test videos. For example, the *Jumping side* testing video contains two cycles of *Jumping side* which are corresponded to the (two) high PSRs. The high PSRs can be further used to accurately *detect* the action occurrences across the test video. Figure 3 (bottom) illustrates some selected frames of *Jumping jack* (left) and *Jumping side* (right) action cycles with their corresponding correlation outputs produced by *Jumping jack* and *Jumping side* filters, respectively. For each frame, its frame number and PSR value are shown. The maximum peak almost occurs at the last frame of each action cycle (temporal domain) upon the location of the actor (spatial domain). A high peak with PSR more than a predefined threshold can be used to accurately detect the action in both spatial and temporal domains.

**Runtime Complexisty and Memory Usage:** The average time for MCCF to classify a $144 \times 180 \times 200$ Weizmann video was 8.15 seconds (real-time) on a Core i7, 3.40 GHz. While, action MACH and [2] required 18.65 seconds and 30 minutes for the same video, respectively. Moreover, most of the other methods in Table 1 used SVM which is shown to be much slower than MCCF [12]. For memory usage, MCCF is very efficient, since the amount of memory required to learn an MCCF is independent of the number of training examples [12]. Whereas, the others suffer from memory overhead, as they need to load all training examples for learning.

| Method | Weizmann | UCF sport |
|---|---|---|
| Huang et al. [19] | 100% | - |
| Cai et al. [21] | 98.7% | 90.6% |
| Wang et al. [17] | 97.8 % | 77.4% |
| Campos et al. [20] | 96.7 % | 80.0% |
| Rodriguez et al. [3] | 86.6% | 69.2% |
| Yeffet & Wolf [22] | - | 79.3% |
| Our method | 97.8% | 82.6% |

**Table 1**. Mean accuracy of our method compared to the state-of-the-art on the Weizmann and UCF sport datasets.

| Normalized intensity | Edge magnitude | Temporal derivative | HoG (5 bins) |
|---|---|---|---|
| 89.4% | 91.2% | 92.3% | 97.8% |

**Table 2**. Mean accuracy of different features (Weizmann)

## 5. CONCLUSION

This paper proposed the application of multi-channel correlation filters for human action detection. The experiments show the competitive performance of our approach against the state-of-the-art with superior computational efficiency. For future work, we will explore MCCFs for action detection.

## 6. REFERENCES


[1] Yang Wang and Greg Mori, "Hidden part models for human action recognition: Probabilistic versus max margin," in *PAMI*, 2011, vol. 33(7), pp. 1310–1323.

[2] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri, "Actions as space-time shapes," in *ICCV*, 2005, pp. 1395–1402.

[3] Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *CVPR*, 2008.

[4] Ivan Laptev, Marcin Marszaek, Cordelia Schmid, and Benjamin Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, 2008, pp. 1–8.

[5] Saad Ali Paul Scovanner and Mubarak Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *ACM Multimedia*, 2007, vol. 1, pp. 357–360.

[6] Alexei A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik, "Recognizing action at a distance," in *ICCV*, 2003, vol. 1, pp. 357–360.

[7] C. F. Hester and D. Casasent, "Multivariant technique for multiclass pattern recognition," *Appl. Opt.*, vol. 19, no. 11, pp. 1758–1761, 1980.

[8] B. V. K. Vijaya Kumar, "Minimum-variance synthetic discriminant functions," *J. Opt. Soc. Am. A*, vol. 3, no. 10, pp. 1579–1584, 1986.

[9] A. Mahalanobis, B. V. K. Vijaya Kumar, and D. Casasent, "Minimum average correlation energy filters," *Appl. Opt.*, vol. 26, no. 17, pp. 3633–3640, 1987.

[10] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *CVPR*, 2010.

[11] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *CVPR*. IEEE, 2005, vol. 1, pp. 886–893.

[12] Hamed Kiani, Terence Sim, and Simon Lucey, "Multi-channel correlation filters," in *ICCV*, 2013.

[13] Saad Ali and Simon Lucey, "Are correlation filters useful for human action recognition?," in *ICPR*, 2010, pp. 2608–2611.

[14] Joseph A. Fernandez and B. V. K. Vijaya Kumar, "Space-time correlation filters for human action detection," in *SPIE*, 2013.

[15] Yicong Tian, Rahul Sukthankar, and Mubarak Shah, "Spatiotemporal deformable part models for action detection," in *CVPR*, 2013, pp. 2642–2649.

[16] BVK Vijaya Kumar, *Correlation pattern recognition*, Cambridge University Press, 2005.

[17] Heng Wang, Muhammad Muneeb Ullah, Alexander Kläser, Ivan Laptev, and Cordelia Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC*, 2009, pp. 1–11.

[18] Alexander Kluser, Marcin Marszalek, and Cordelia Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC*, 2008, pp. 1–10.

[19] Zhi Feng Huang, Weilong Yang, Yang Wang, and Greg Mori, "Latent boosting for action recognition," in *BMVC*, 2011, pp. 1–11.

[20] Teofilo de Campos, Mark Barnard, Krystian Mikolajczyk, Josef Kittler, Fei Yan, William J. Christmas, and David Windridge, "An evaluation of bags-of-words and spatio-temporal shapes for action recognition," in *WACV*, 2011, pp. 344–351.

[21] Qiao Cai, Yafeng Yin, and Hong Man, "Dspm: Dynamic structure preserving map for action recognition," in *ICME*, 2013, pp. 1–6.

[22] Lahav Yeffet and Lior Wolf, "Local trinary patterns for human action recognition," in *ICCV*, 2009, pp. 492–497.