Encoding Scale into Fisher Vector for Human Action Recognition

Bowen Zhang [#§], Hanli Wang [#§*]

[#] *Department of Computer Science and Technology, Tongji University, Shanghai, China*
[§] *Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai, China*
1023zhangbowen@tongji.edu.cn, hanliwang@tongji.edu.cn

***Abstract*—In this paper, a new kind of Fisher Vector (FV) model, named Scale FV (ScaleFV), is proposed to ameliorate visual feature encoding for human action recognition. Although several researches have been proposed for feature encoding, the temporal scale information is almost ignored. Similar to the spatial scale information which has shown to be important in extracting and encoding visual features, the temporal scale information also plays an important role in video content analysis based on our investigation. To demonstrate this, a definition of temporal scale in videos is given, and it is presented that both of the spatial and temporal scale information can be encoded into the FV model by slightly modifying the underlying Gaussian Mixture Models (GMM). Furthermore, an enhanced FV model termed as Combined FV (CombFV) is designed to capture both position and scale information for human action recognition. Comparative experiments are carried out to demonstrate the superior performance of the proposed methods.**

***Index Terms*—Human action recognition, Gaussian Mixture Model, Fisher Vector, temporal scale, spatial scale.**

I. Introduction

Human action recognition is a challenging problem in computer vision. In this work, in order to tackle the problem about how to automatically recognize human actions in videos, a novel kind of feature encoding model, called Scale Fisher Vector (ScaleFV), is proposed which combines the discrimination power of Fisher Vector (FV) [1] and scale information. In the literature, a number of works have been proposed for human action recognition, such as [2], [3], [4], [5] to name a few, most of which are based on the Bag of Words (BoW) paradigm [6]. With BoW, keypoints of videos are extracted by feature extraction techniques followed by feature description. Then, feature encoding is employed to represent each video by a feature vector for subsequent recognition.

[*]Corresponding author (H. Wang). This work was supported in part by the National Natural Science Foundation of China under Grant 61472281, the "Shu Guang" project of Shanghai Municipal Education Commission and Shanghai Education Development Foundation under Grant 12SG23, the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning (No. GZ2015005), and the Fundamental Research Funds for the Central Universities under Grant 0800219270.

For feature extraction and description, it is shown in [7], [8] on how to exploit spatial information at the pixel level. To utilize temporal information, the Dense Trajectory (DT) method [4] is proposed to use dense sampling to capture keypoints and the Farnebäck optical flow [9] is employed to track keypoints in continuous frames to generate trajectory based features. It has been shown in [4] that DT outperforms the KLT trajectory [10] and SIFT trajectory [11]. Later, Wang *et al.* further propose an Improved DT (IDT) method [12] which achieves state-of-the-art results in a number of benchmark challenging datasets. Due to its competitive performance, IDT is used in this work to extract and describe keypoints.

Regarding feature encoding, hard quantization with clustering is generally employed owing to its effectiveness and efficiency. However, hard quantization only includes frequency information and does not take the information like mean and covariance into account, which may hinder the performance. Furthermore, in order to achieve good performance, hard quantization usually requires to enlarge the number of clusters, leading to lots of computations consumed. To tackle the shortcomings of hard quantization, the FV model [1] is designed. With the aid of Gaussian Mixture Model (GMM), FV can generate a high dimension of vector by clustering a small number of Gaussian models. However, the original FV model lacks spatial-temporal information and scale information. To add spatial-temporal information, the Spatial FV (SFV) [13] and Spatial-Temporal FV (STFV) [14] are proposed to encode FV by incorporating spatial-temporal information into GMM. Although SFV and STFV show superior performances than FV, the performance improvement is marginal. In order to fully leverage the discriminative ability of FV, scale information should be considered. However, the embedding of scale information into FV is not well studied for videos.

In this work, the spatial-temporal scale information is introduced into FV for human action recognition, with the following two major contributions. First, the proposed ScaleFV considers the spatial-temporal scale information into FV to boost human action recognition. Second, a more comprehensive FV model called CombFV is further designed which considers both scale information and spatial-temporal position information. The proposed ScaleFV, CombFV and a number of combinations of ScaleFV, CombFV and STFV are evaluated on two challenging benchmark datasets including Hollywood-2 [15] and HMDB51 [16]. The rest of this paper is organized

as follows. The details of the proposed ScaleFV, CombFV and the related fusion are introduced in Section II. The experimental results are presented in Section III. Finally, Section IV concludes this paper.

II. Proposed Scale Feature Vector Models

A. Temporal Scale in Videos

Although spatial scale has been studied for visual signals such as [17], the temporal scale information is not well investigated. This may be due to the fact that temporal scale is only meaningful in videos. Unlike the clear definition of spatial scale in images, it is not a trivial task to define temporal scale in videos. Intuitively, temporal scale means the time period of a motion occupied from the beginning to the end. However, it is hard to automatically figure out when a motion starts and ends in a video. To address this issue, we define the temporal scale in videos as the length of an event trajectory within a specific time period. For instance, when using the IDT method [12] to extract features in a given video, the length of each trajectory can be produced, which is utilized to represent the temporal scale of the corresponding feature. As shown in Fig. 1, different motions (such as 'Eat', 'Run' and 'SitDown') have different distributions in frequency of trajectory length. These different distributions can be used to improve the discrimination ability for human action recognition.

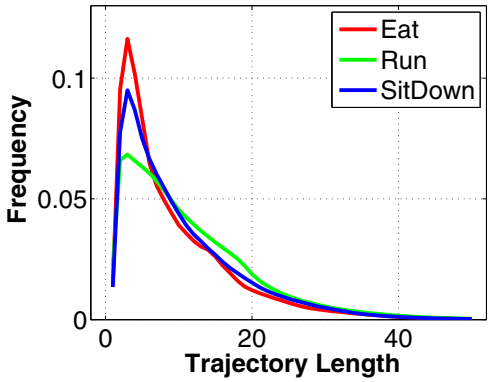

Fig. 1. Frequency of trajectory length about three action classes ('Eat', 'Run' and 'SitDown') in Hollywood-2.

B. Fisher Vector

Before introducing the proposed scale FV models, a brief introduction to FV is presented. Assume there are M keypoints which are extracted from the training set $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_M\}$. These keypoints are then described by a description algorithm as $\mathbf{y}_t = f(\mathbf{x}_t), t = 1, \cdots, M$, where $f(\cdot)$ is a description function. Then, a GMM p is used to cluster these feature vectors $\mathbf{y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_M\}$. We define $\theta = \{\pi_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$, where π_i, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ stand for the weight, mean and covariance of the i^{th} Gaussian model, respectively, the GMM can be formulated as

$$p(\mathbf{y}_t; \theta) = \sum_{i=1}^{K} \pi_i N(\mathbf{y}_t; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{1}$$

where K represents the number of Gaussian models, $N(\cdot)$ is the multidimensional Gaussian density function, and the weight π_i is calculated as [13]: $\pi_i = \exp(\alpha_i)/\sum_j \exp(\alpha_j)$, where α_i satisfies the requirement that $\pi_i \geq 0$ and $\sum_{i=1}^{K} \pi_i = 1$. In general, GMM can be generated by the Expectation Maximization (EM) algorithm. Based on the Bayesian theorem, the posterior possibility can be defined as

$$q_i(t) = \frac{\pi_i N(\mathbf{y}_t; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{j=1}^{K} \pi_j N(\mathbf{y}_t; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \tag{2}$$

Accordingly, FV can be produced with the following three formulae by derivation operations.

$$\frac{\partial \ln p(\mathbf{y}_t)}{\partial \alpha_i} = q_i(t) - \pi_i, \tag{3}$$

$$\frac{\partial \ln p(\mathbf{y}_t)}{\partial \boldsymbol{\mu}_i} = q_i(t)\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_t - \boldsymbol{\mu}_i), \tag{4}$$

$$\frac{\partial \ln p(\mathbf{y}_t)}{\partial \boldsymbol{\Sigma}_i^{-1}} = q_i(t)\frac{\boldsymbol{\Sigma}_i - diag((\mathbf{y}_t - \boldsymbol{\mu}_i)^2)}{2}, \tag{5}$$

where $(\mathbf{y}_t - \boldsymbol{\mu}_i)^2$ in Eq. (5) is the element-wise square of $(\mathbf{y}_t - \boldsymbol{\mu}_i)$, and $diag(\boldsymbol{x})$ means using a diagonal matrix to represent vector $\boldsymbol{x}$, where $diag(\boldsymbol{x})(ind, ind) = \boldsymbol{x}(ind)$ with ind standing for the index in $\boldsymbol{x}$. It should be noticed that diagonal elements are used in Eq. (5) to represent gradients. The final FV representation for a video is to average all features' gradients. Therefore, the dimension of FV is $(2D+1)K$, where D is the dimension of the feature vector $\mathbf{y}_t$.

C. Scale Fisher Vector

Based on the aforementioned introduction to FV, the proposed Scale FV (ScaleFV) model is detailed as follows. As compared with the traditional FV model, ScaleFV takes the scale information $\mathbf{s} = \{\sigma_t, \tau_t, t = 1, \cdots, M\}$ into consideration, where σ_t and τ_t are the spatial scale and temporal scale of the keypoint $\mathbf{x}_t$, respectively. Inspired by [13], we modify the original GMM in Eq. (1) by multiplying scale GMM as

$$p(\mathbf{y}_t, \mathbf{s}_t; \theta, \epsilon) = \sum_{i=1}^{K} \pi_i N(\mathbf{y}_t; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \sum_{j=1}^{J} \lambda_{ji} N(\mathbf{s}_t; \boldsymbol{\delta}_{ji}, \boldsymbol{Z}_{ji}), \tag{6}$$

where λ_{ji}, $\boldsymbol{\delta}_{ji}$ and $\boldsymbol{Z}_{ji}$ stand for the weight, mean and covariance of the j^{th} scale Gaussian model, and J is the number of scale GMM models. In a similar manner, we define $\lambda_{ji} = \exp(\alpha_{ji})/\sum_m \exp(\alpha_{mi})$, where α_{mi} satisfies the requirement that $\lambda_{ji} \geq 0$ and $\sum_{j=1}^{J} \lambda_{ji} = 1$.

The proposed ScaleFV is defined by $\frac{\partial \ln p(\mathbf{y})}{\partial \alpha_i}$, $\frac{\partial \ln p(\mathbf{y})}{\partial \boldsymbol{\mu}_i}$, $\frac{\partial \ln p(\mathbf{y})}{\partial \boldsymbol{\Sigma}_i^{-1}}$, $\frac{\partial \ln p(\mathbf{y})}{\partial \alpha_{ji}}$, $\frac{\partial \ln p(\mathbf{y})}{\partial \boldsymbol{\delta}_{ji}}$ and $\frac{\partial \ln p(\mathbf{y})}{\partial \boldsymbol{Z}_{ji}^{-1}}$. The first three terms are the same as that defined in Eqs. (3, 4, 5). Regarding the last three terms related to the proposed scale GMM, after mathematical manipulations they are formulated as

$$\frac{\partial \ln p(\mathbf{y}_t, \mathbf{s}_t)}{\partial \alpha_{ji}} = q_i(t)(r_{ji}(t) - \lambda_{ji}), \tag{7}$$

$$\frac{\partial \ln p(\mathbf{y}_t, \mathbf{s}_t)}{\partial \boldsymbol{\delta}_{ji}} = q_i(t) r_{ji}(t) \boldsymbol{Z}_{ji}^{-1}(\mathbf{s}_t - \boldsymbol{\delta}_{ji}), \tag{8}$$

$$\frac{\partial \ln p(\mathbf{y}_t, \mathbf{s}_t)}{\partial \boldsymbol{Z}_{ji}^{-1}} = q_i(t) r_{ji}(t) \frac{\boldsymbol{Z}_{ji} - diag((\mathbf{s}_t - \boldsymbol{\delta}_{ji})^2)}{2}, \quad (9)$$

where

$$q_i(t) = \frac{\pi_i N(\mathbf{y}_t; \mu_i, \boldsymbol{\Sigma}_i) \sum_{j=1}^{J} \lambda_{ji} N(\mathbf{s}_t; \boldsymbol{\delta}_{ji}, \boldsymbol{Z}_{ji})}{\sum_{l=1}^{K} \pi_l N(\mathbf{y}_t; \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) \sum_{j=1}^{J} \lambda_{jl} N(\mathbf{s}_t; \boldsymbol{\delta}_{jl}, \boldsymbol{Z}_{jl})}, \quad (10)$$

$$r_{ji}(t) = \frac{\lambda_{ji} N(\mathbf{s}_t; \boldsymbol{\delta}_{ji}, \boldsymbol{Z}_{ji})}{\sum_{l=1}^{J} \lambda_{li} N(\mathbf{s}_t; \boldsymbol{\delta}_{li}, \boldsymbol{Z}_{li})}. \quad (11)$$

As a consequence, the dimension of the proposed ScaleFV is $(2D + 1 + J(2d + 1))K$, where $d = 2$ indicates the dimension of the scale information. We can see SFV [13] and STFV [14] are modifications of ScaleFV by changing the scale information $\mathbf{s}$ with the spatial position information $\mathbf{l} = \{\alpha_t, \beta_t\}$ and the spatial-temporal position information $\mathbf{l} = \{\alpha_t, \beta_t, \gamma_t\}$, respectively, where α_t, β_t and γ_t are the spatially horizontal, spatially vertical and temporal position of $\mathbf{x}_t$ in a video.

D. Combined Fisher Vector

The proposed Combined FV (CombFV) encodes both spatial-temporal position and scale information into FV. In CombFV, the scale information $\mathbf{s}$ is extended to $\mathbf{u} = \{\alpha_t, \beta_t, \gamma_t, \sigma_t, \tau_t\}$, which contains both spatial-temporal position and scale information of keypoints. Therefore, the GMM of CombFV is updated as

$$p(\mathbf{y}_t, \mathbf{u}_t; \theta, \epsilon) = \sum_{i=1}^{K} \pi_i N(\mathbf{y}_t; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \sum_{j=1}^{J} \lambda_{ji} N(\mathbf{u}_t; \boldsymbol{\delta}_{ji}, \boldsymbol{Z}_{ji}). \quad (12)$$

The representations of CombFV remain the same as Eqs. (3,4,5,7,8,9) but using $\mathbf{u}_t$ instead of $\mathbf{s}_t$, and the dimension of CombFV is updated as $(2D + 1 + J(2 \times 5 + 1))K$.

E. Fusion of Fisher Vector Models

In order to further encode scale and spatial-temporal position information, we employ early fusion to combine ScaleFV, STFV and CombFV, so that different types of FVs are concatenated to generate a new vector. With this fusion, each FV model has its own emphasis. The proposed ScaleFV focuses on capturing scale information while STFV emphasizes on spatial-temporal position information, and the proposed CombFV grasps the relation between scale and spatial-temporal position information. By this kind of fusion, the scale and position information can be better utilized for performance enhancement. As will be presented in the following experiments, the two types of fusion including STFV+CombFV and STFV+CombFV+ScaleFV are able to truly improve the human action recognition performance.

III. Experimental Results

In order to evaluate the proposed ScaleFV, CombFV and the extended fusion models, two benchmark video datasets Hollywood-2 [15] and HMDB51 [16] are used for comparative experiments. The Hollywood-2 dataset possesses 823 training videos and 884 testing videos with 12 human action classes. We use the recommendation setup for the separation of training and testing videos. The criterion of mean Average Precision (mAP) is used to present the recognition performance for Hollywood-2. Regarding HMDB51, it contains 6,766 videos from 51 classes and is featured for its diversity of motions and complex backgrounds. The HMDB51 videos are divided into three different splits and we adopt the standard setup to separate training and testing videos. About the recognition criterion, the average accuracy is employed for HMDB51.

In our implementation, the IDT method [12] without human detection is applied to extract and describe keypoints, and four popular descriptors are used including trajectory, histogram of oriented gradient, histogram of optical flow and motion boundary histogram. Moreover, since it is time-consuming to calculate different scale GMMs for each cluster, we generate one scale GMM instead as

$$p(\mathbf{y}_t, \mathbf{s}_t; \theta, \epsilon) = \sum_{i=1}^{K} \pi_i N(\mathbf{y}_t; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \sum_{j=1}^{J} \lambda_j N(\mathbf{s}_t; \boldsymbol{\delta}_j, \boldsymbol{Z}_j). \quad (13)$$

The difference between Eq. (6) and Eq. (13) is that the index in Eq. (13) is j instead of ji. Under this approximation, for one descriptor, we generate only one GMM for scale information.

For training our models, a total of $256,000$ features are sampled from training videos and the Principle Component Analysis is further used to reduce the feature dimension to half. Then, these features are applied to generate GMMs and scale GMMs. In the experiments, the number of clusters K for GMM is set to 256. The FVs for different descriptors are concatenated for subsequent classification with the linear Support Vector Machine employed.

The comparative results are summarized in Table I with different settings of the scale GMM cluster number J. The detailed analyses of these results are presented as follows.

TABLE I
COMPARATIVE RESULTS OF STFV, SCALEFV, COMBFV AND THEIR COMBINATIONS ON HOLLYWOOD-2 AND HMDB51.

Methods	J	Hollywood-2	HMDB51
STFV [14]	1	65.92%	60.39%
	2	65.92%	60.35%
	3	66.52%	61.07%
	4	65.69%	60.61%
	5	65.80%	60.96%
ScaleFV	1	64.69%	58.36%
	2	64.84%	58.65%
	3	64.79%	58.61%
	4	64.61%	58.41%
	5	64.65%	57.87%
CombFV	1	66.28%	60.41%
	2	66.29%	60.50%
	3	66.48%	60.41%
	4	66.53%	60.28%
	5	66.50%	60.56%
STFV+CombFV	-	66.72%	**61.50%**
STFV+CombFV+ScaleFV	-	**66.96%**	61.07%

1) STFV and ScaleFV: As shown in Table I, STFV [14] outperforms ScaleFV on Hollywood-2 and HMDB51. This

TABLE II
COMPARISON BETWEEN STATE-OF-THE-ART APPROACHES AND THE PROPOSED STFV+COMBFV AS WELL AS STFV+COMBFV+SCALEFV.

	Hollywood-2	HMDB51
Jain *et al.* [18]	62.5%	52.1%
Oneata *et al.* [14]	63.3%	54.8%
Wang *et al.* [12]	64.3%	57.2%
Wu *et al.* [19]	64.5%	-
Simonyan *et al.* [20]	-	59.4%
STFV+CombFV	66.72%	**61.50%**
STFV+CombFV+ScaleFV	**66.96%**	61.07%

may be due to the reason that STFV owns a larger dimension than ScaleFV, *i.e.*, the dimension of STFV is $(2D+1+7J)K$, while the dimension of ScaleFV is $(2D+1+5J)K$. Such a comparison shows that it may not be sufficient to consider the scale information solely for human action recognition, which inspires us to consider a more powerful encoding method to take both scale and spatial-temporal position information into account.

2) CombFV and STFV: It can be concluded that CombFV outperforms STFV by 0.4% in average on Hollywood-2 with the same cluster number J. This indicates that the scale and position information can complement each other. However, on HMDB51, the performances of CombFV and STFV are similar to each other.

3) STFV+CombFV and STFV+CombFV+ScaleFV: As described in Section II-E, early fusion can be applied to combine different types of FVs for performance improvement. In the current work, we consider two scenarios of combination, including STFV+CombFV and STFV+CombFV+ScaleFV. As shown in Table I, STFV+CombFV outperforms ScaleFV, STFV and CombFV, which reveals that the fusion of STFV and CombFV really works. Moreover, STFV+CombFV+ScaleFV further improves the results on Hollywood-2 while slightly declining the performance on HMDB51, which is due to the fact that the distributions of the scale information in HMDB51 are similar so that the performance achieved by fusion with ScaleFV is not useful. Note that the numbers of scale GMMs J are empirically set to 3, 5 and 2 for STFV, CombFV and ScaleFV, respectively.

4) Comparison with State-of-the-Arts: In addition, we make a comparison with a number of state-of-the-art approaches in the literature, including the divergence-curl-shear descriptor based approach [18], the spatial-temporal pyramid based approach [14], the traditional FV approach [12] and the deep neural network based approaches [19], [20]. The summary of the comparison is presented in Table II, where it can be clearly seen that the proposed approaches including STFV+CombFV and STFV+CombFV+ScaleFV are superior to the comparative approaches on the datasets of Hollywood-2 and HMDB51. In fact, the proposed scale FV models can also be applied to these state-of-the-art approaches to possibly boost human action recognition performances, which will be open for future investigation.

5) Complexity: With the help of multi-thread techniques, the computations of ScaleFV and CombFV are efficient, *e.g.*, the consuming-time of GMM on scale information is limited (≤ 1 minute on a computer with CPU of Corei7 3.6GHz).

IV. CONCLUSION

In this paper, a novel feature encoding model called ScaleFV is designed, which is further improved to produce the CombFV model by combining both the scale and spatial-temporal position information. Based on these two methods, several early fusion methods of FV models are explored. The effectiveness of the proposed methods is verified on the benchmark Hollywood-2 and HMDB51 datasets. The comparative experimental results demonstrate that the combination of scale and spatial-temporal position information is able to improve the performance for human action recognition and outstrips a number of state-of-the-art approaches.

REFERENCES

[1] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *ECCV'10*, Sept. 2010, pp. 143–156.

[2] I. Laptev, "On space-time interest points," *IJCV*, vol. 64, no. 2-3, pp. 107–123, Sept. 2005.

[3] L. Wang, Y. Qiao, and X. Tang, "Latent hierarchical model of temporal structure for complex activity classification," *TIP*, vol. 23, no. 2, pp. 810–822, Feb. 2014.

[4] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *IJCV*, vol. 103, no. 1, pp. 60–79, May 2013.

[5] L. Wang, Y. Qiao, and X. Tang, "Motionlets: Mid-level 3D parts for human motion recognition," in *CVPR'13*, Jun. 2013, pp. 2674–2681.

[6] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *ICCV'03*, Oct. 2003, pp. 1470–1477.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[8] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *ECCV'06*, May 2006, pp. 404–417.

[9] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *SCIA'03*, Jun. 2003, pp. 363–370.

[10] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI'81*, Aug. 1981, pp. 674–679.

[11] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," in *CVPR'09*, Jun. 2009, pp. 2004–2011.

[12] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *ICCV'13*, Dec. 2013, pp. 3551–3558.

[13] J. Krapac, J. Verbeek, and F. Jurie, "Modeling spatial layout with fisher vectors for image categorization," in *ICCV'11*, Nov. 2011, pp. 1487–1494.

[14] D. Oneata, J. Verbeek, and C. Schmid, "Action and event recognition with fisher vectors on a compact feature set," in *ICCV'13*, Dec. 2013, pp. 1817–1824.

[15] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in *CVPR'09*, Jun. 2009, pp. 2929–2936.

[16] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *ICCV'11*, Nov. 2011, pp. 2556–2563.

[17] T. Lindeberg, *Scale-space theory in computer vision*. Springer Science & Business Media, 1993.

[18] M. Jain, H. Jégou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *CVPR'13*, Jun. 2013, pp. 2555–2562.

[19] Z. Wu, Y.-G. Jiang, J. Wang, J. Pu, and X. Xue, "Exploring inter-feature and inter-class relationships with deep neural networks for video classification," in *ACM MM'14*, Nov. 2014, pp. 167–176.

[20] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS'14*, Dec. 2014, pp. 568–576.