Vision-based Human Action Recognition on Pre-trained AlexNet

NorulNarijah Mohamed Zamri
Fac. of Info. Science &Tech.
Multimedia University, Malaysia
1121117295@student.mmu.edu.my

Goh Fan Ling
DRSOFT Sdn Bhd
Melaka, Malaysia
stella.goh.fan.ling.88@gmail.com

Pang Ying Han*
Fac. of Info. Science &Tech.
Multimedia University, Malaysia
yhpang@mmu.edu.my

Ooi Shih Yin
Fac. of Info. Science &Tech.
Multimedia University, Malaysia
syooi@mmu.edu.my

***Abstract*—The Deep learning analysis has been extensively carried out in the context of object/ pattern recognition due to its excellence in feature extraction and classification. However, the superior performance just can be guaranteed with the availability of huge amounts of training data and also high-specification data processing unit to process the data deeper at high speeds. Hence, another alternative is by applying transfer learning. In transfer learning, a neural network model is first trained on a data similar to the targeted data. With that, the knowledge such as features, weights etc. could be leveraged from the trained model to train the new model. In this project, a vision-based human action recognition via a transfer learning is conducted. Specifically, in the proposed approach, the earlier layers of a pre-trained AlexNet is preserved since those extracted low-level features are characterizing generic features which are common to most data. However, the pre-train network is fine-tuned based on our interested data, that is human action data. Since AlexNet requires input data of size 227*227*3, the frames of each video are processed into 3 different templates. The three computed templates are: (1) Motion History Image carrying spatio-temporal information, (2) Binary Motion Energy Image incorporating motion region information and (3) optical flow template holding accumulative motion speed information. The proposed approach is validated on two publicly available databases, which are Weizmann database and KTH database. From the empirical results, a promising performance is obtained with about 90% accuracy from the databases.**

Keywords-human action; pre-trained AlexNet; binary energy image; motion historical image; optical flow magniture

I. Introduction

Recently, the research on human action recognition is conducted extensively [1][2][3]. The key reason of the popularity of this research area is that human action recognition endows numerous real world applications: vision-based surveillance, gaming and entertainment, human-machine interaction etc.

Basically, the existing state-of-the-art approaches of human action recognition could be classified into two spheres: shallow approach and deep learning approach. HOG (Histogram of Oriented Gradient)-3D [4] and HOF (Histogram of Optical Flow) [5] as well as Localized Temporal Representation [6] are the examples of shallow approach. These handcrafted methods require domain expert's input or prior knowledge for designing optimal descriptors for feature extraction and representation. On the other hand, feature learning based on deep architecture has been gaining increasing interest from kinds of computer vision domains due to its powerful feature learning. The examples of deep learning methods include convolutional neural network [7], two-stream ConvNet [8], 3D-CNN with 3D motion cuboid [9], deep nets [10] etc. The downsides of deep learning approach are huge amount of data, at least thousands of labeled samples, is required, expensive computation that could take several weeks to train completely from scratch.

Understanding that human is able to learn thousands of objects from a few samples. The reason for this is that human accumulates knowledge in the time and transfers the information for learning new objects. Inspired with this, researchers are confident to the possibility of transferring the deep learning models that are trained on a specific data to a new data set [11]. In other words, a pre-trained neural network model is possible to be tuned to handle a new duty. This is known as transfer learning.

There are two means of applying transfer learning with deep networks: (1) utilizing pre-trained network and preserve the learned weights of all layers, except the last three layers, for feature representation, then a generic classifier is performed; (2) the pre-trained network is preserved and network weights are updated through training with a new set of data [12]. Our proposed approach corresponds to the second type.

Although AlexNet is not specifically designed for non-image data, this kind of convolutional neural network attains state-of-the-art performance on problems like document classification. Hence, there is possible to try such convolutional neural network on time series or sequence input data. In view of this, we explore the feasibility of transfer learning a pre-trained network, specifically AlexNet, in the context of action recognition in our project. AlexNet model is a deep convolutional neural network (CNN) and millions of labelled images from the ImageNet Large-Scale Visual

978-1-7281-3500-7/19/$31.00 ©2019 IEEE

Recognition Challenge dataset [13][14] are deployed to train the network. In the network architecture, there are 5 convolutional layers, 2 normalization layers, 3 max pooling layers, 3 fully connected layers and a linear layer with Softmax activation in the output. The architecture of AlexNet is demonstrated in Fig. 1.

In this approach, the earlier layers of AlexNet is preserved since those layers unearth those generic and low-level features. These generic features include edges, blobs etc. and they are typically invariant across data [15]. On the other hand, the pre-trained network is tuned through substituting the last 3 layers with fully connected, Softmax and classification output layers according to our new data.

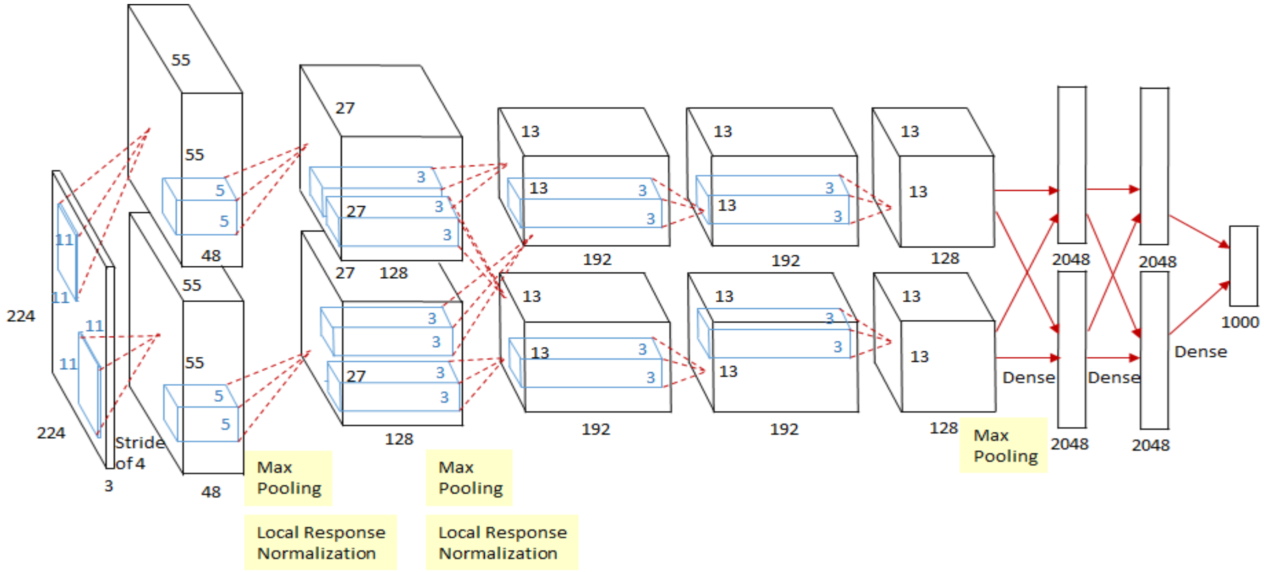


Figure 1. The architecture of AlexNet model (source of the figure: https://medium.com/coinmonks/paper-review-of-alexnet-caffenet-winner-in-ilsvrc-2012-image-classification-b93598314160).

II. Methodology

Directly feeding video data into AlexNet might not be appropriate since AlexNet is not a sequential network. On top of this, AlexNet requires input data of size 227*227*3. Hence, video data should be processed before inputting into the pre-trained AlexNet. Video data is a series of frames. In this project, the video frames are processed into three different templates carrying (1) **spatio-temporal** information via Motion History Image, (2) **motion region** information via Binary Motion Energy and (3) distribution of **apparent motion** via Optical Flow estimation. Specifically, Motion History Image method is used to characterize the spatio-temporal information of an action; Binary Motion Energy Image method is employed to encode the temporal information; for optical flow estimation, Lucas-Kanade method is implemented for the distribution of apparent motion.

A. Motion History Image (MHI)

MHI transforms a 3-dimensional video data into a two-dimensional form. Through this process, the recency of spatio-temporal change of a motion could be encoded. MHI template is derived as follow [16]:

$$H_\tau(x,y,t) = \begin{cases} \tau, & if D(x,y,t) = 1 \\ \max\{0, H_\tau(x,y,t-1) - 1\}, & otherwise \end{cases} \qquad (1)$$

$D(x,y,t)$ is motion estimation function at location $D(x,y)$ at time t. τ defines the motion temporal information. Sample of MHI template is illustrated in Fig. 2. The darker regions indicate the earlier happened motion, and the brighter areas imply the latter action.

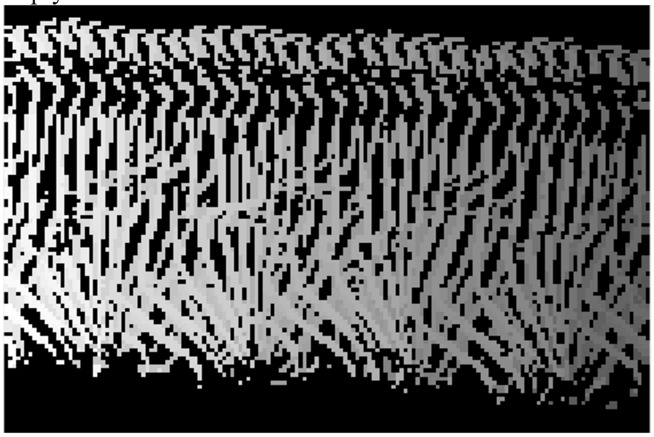


Figure 2. MHI sample of walking motion

B. Motion Energy Image (MEI)

MEI symbolizes the regions of motion. In other words, it locates where motion happened in a single binary template. MEI can be computed through the equation below:

$$E_\tau(x,y,t) = \bigcup_{i=0}^{\tau-1} D(x,y,t-i) \qquad (2)$$

$D(x, y, t)$ is the binary data representing the motion regions. Fig. 3 illustrates MEI sample of bending motion. From the figure, we can see that the area of the motion. Utilizing such information, the movement as well as the viewing condition (in term of angle) could be determined [17].

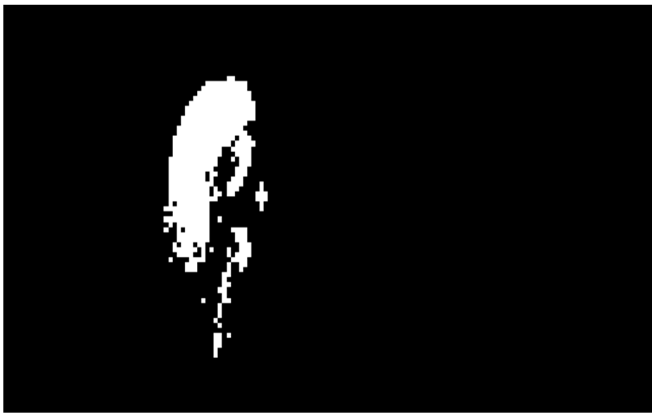


Figure 3. MEI sample of bending motion

C. Optical Flow

Optical flow representation portrays the velocity of every single pixel point in the image/frame. There are many algorithms for computing the optical flow [18][19]. Among these, Lucas-Kaneda algorithm is one of the most popular methods for its computational efficiency [20]. With hypothesis of constant brightness, it is assumed that the optical flow is hold for all pixels at p location. Image flow vector (V_x, V_y) satisfies:

$$I_x(q1)V_x + I_y(q1)V_y = -I_t(q1)$$
$$I_x(q2)V_x + I_y(q2)V_y = -I_t(q2)$$
$$\vdots$$
$$I_x(qn)V_x + I_y(qn)V_y = -I_t(qn) \qquad (3)$$

$q1, q2, \dots, qn$ pixels inside a window and $I_x(qi)$, $I_y(qi)$ and $I_t(qi)$ are the partial derivatives of image I at x, y position and time t, evaluated at the point qi and at the current time.

The above equation can be expressed as $Av = b$:

$$A = \begin{bmatrix} I_x(q1) & I_y(q1) \\ \vdots & \vdots \\ I_x(qn) & I_y(qn) \end{bmatrix}$$

$$v = \begin{bmatrix} V_x \\ V_y \end{bmatrix}$$

$$b = \begin{bmatrix} -I_t(q1) \\ \vdots \\ -I_t(qn) \end{bmatrix}$$

By the least squares principle, we have:

$$A^T A v = A^T b$$
$$v = (A^T A)^{-1} A^T b \qquad (4)$$

A^T is the transpose of A. Hence, the optical flow velocity could be computed as follow (2x2 system is shown as example):

$$\begin{bmatrix} V_x \\ V_y \end{bmatrix} = \begin{bmatrix} \sum_i I_x(qi)^2 & \sum_i I_x(qi) I_y(qi) \\ \sum_i I_y(qi) I_x(qi) & \sum_i I_y(qi)^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_i I_x(qi) I_t(qi) \\ -\sum_i I_y(qi) I_t(qi) \end{bmatrix} \qquad (5)$$

where the central matrix is an inverse matrix.

The motion speed at location (x,y) in a frame:

$$s(x, y) = \sqrt{{V_x}^2 + {V_y}^2} \qquad (6)$$

An accumulative motion speed of each pixel location is adopted for optical flow template computation. Fig. 4 illustrate the accumulative motion speed across the video frame.

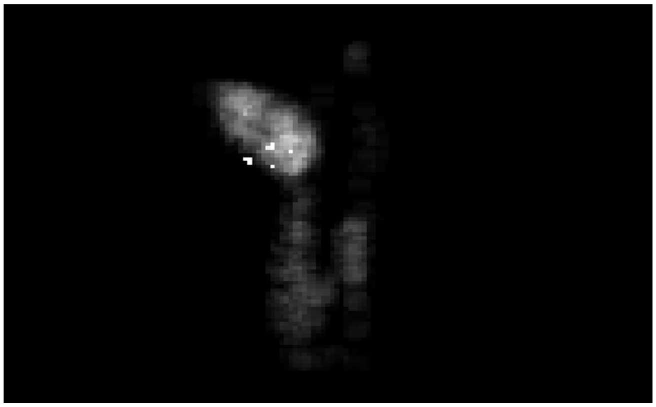


Figure 4. Accumulative motion speed of boxing motion

III. Results and Discussions

A. Datasets

In this paper, two public available human action databases are adopted for performance evaluation: (1) KTH database, and (2) Weizmann database. In KTH database, there are six classes of human actions (i.e. walking, hand waving, boxing, jogging, hand clapping and running) performed by 25 users in scenarios: (1) outdoors, (2) outdoors with scale variation, (3) outdoors with different clothes and (4) indoors. Samples of the data are illustrated in Fig. 5.

Weizmann database contains 90 videos which are done by 9 users. 10 action classes include: (1) walk, (2) jumping-jack, (3) run, (4) jump-in-place-on-two-legs, (5) skip, (6) gallop sideways, (7) wave-one-hand, (8) wave-two hands, (9) jump-forward-on-two-legs, and (10) bend. Samples of action are shown in Fig. 6.

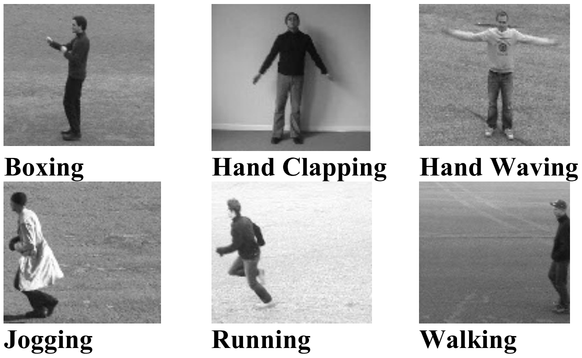

Boxing **Hand Clapping** **Hand Waving**
Jogging **Running** **Walking**

Figure 5. KTH database.

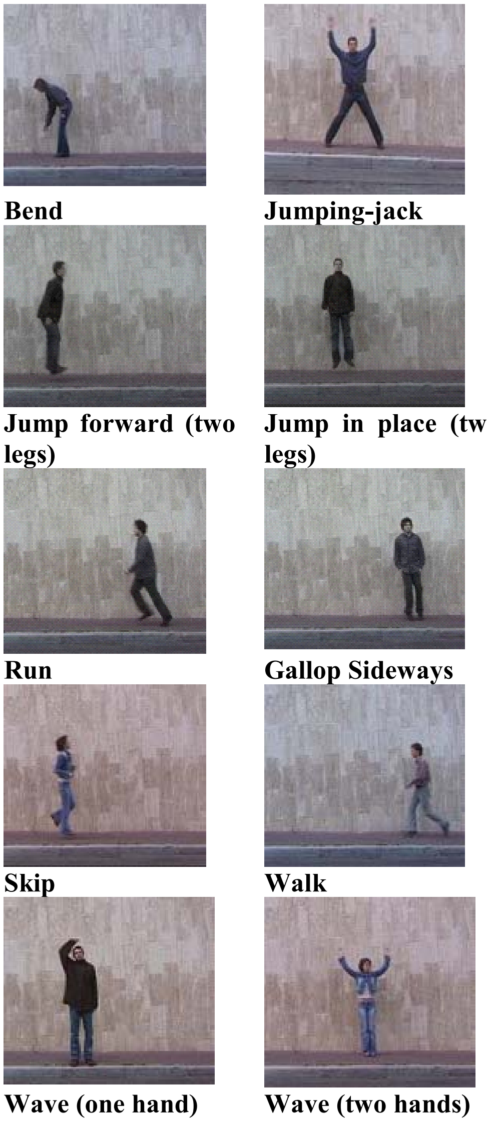
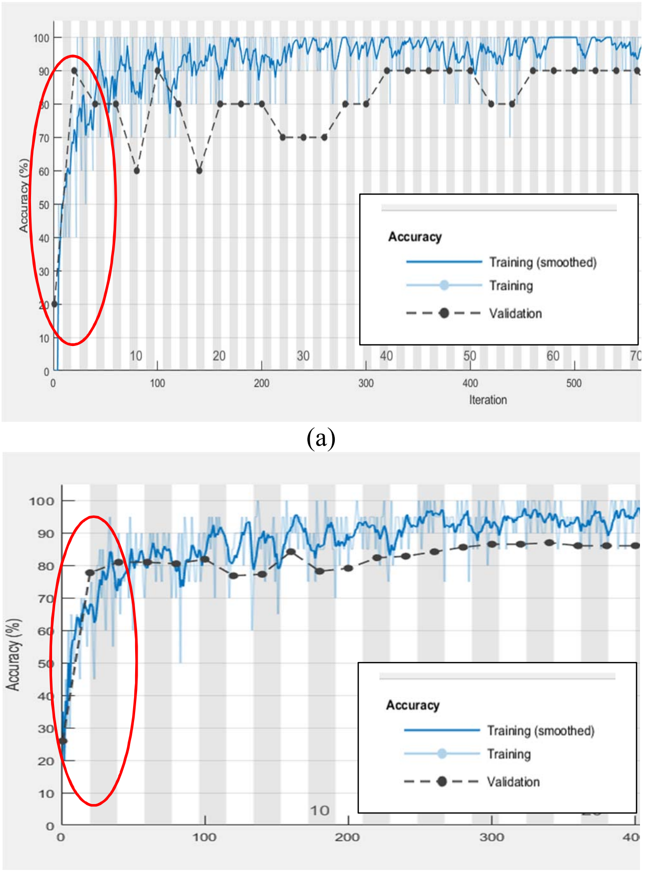

Bend **Jumping-jack**
Jump forward (two legs) **Jump in place (two legs)**
Run **Gallop Sideways**
Skip **Walk**
Wave (one hand) **Wave (two hands)**

Figure 6. Weizmann database.

B. *System Validation Performance*

In this experiment, both Weizmann dataset and KTH database are employed. In Weizmann dataset, leave-one-out testing protocol is adopted, where ONE subject is randomly selected as testing set, the remaining subjects' videos are for training. On the other hand, in KTH database, videos from 16 subjects are used for training and the remaining subjects' videos are as testing sets. Fig. 7 illustrates the training and validation accuracies of the system. Since the front layers (except the last three layers) are pre-trained, the training and validation accuracies feasibly escalate faster at the early epochs (red circled region). Such a promising performance obtained in the starting epochs is due to the initial pre-trained network layers have been trained to learn low-level features from the training data of AlexNet [12]. Examples of these features include blobs, edges, etc.



(a)

(b)

Figure 7. Screenshot of the training process for AlexNet on (a) Weizmann database and (b) KTH database. Validation accuracy grows faster at the early epochs.

From Fig. 7, we also can observe that the validation accuracy increases promptly in the earlier iterations but shows slower pace of increment afterwards. Anyhow, it is observed that there is an improvement in performance when the iteration number increases from 200 to 350. This shows that the network is robust to overfitting due to the data augmentation.

Table I and II records the validation performance of different iteration numbers with batch size = 10 and 20 for

Weizmann database and KTH database, respectively. From Table I, we can observe that the performance is improving along with the number of iteration. However, the performance shows degradation after a certain point, such as at 1600 iterations in the case of batch size=10 and at 800 iterations when batch size = 20. This is because training for too many iterations may result in overtraining/ overfitting problem. We also can observe the similar finding from Table II.

TABLE I. VALIDATION PERFORMANCE OF DIFFERENT ITERATION NUMBERS WITH BATCH SIZE = 10 AND 20 ON WEIZMANN DATABASE

Iterations	**Accuracy (%)**	
	Batch size	
	10	20
100	80*	70
200	80	80
400	90	90
800	90	70
1600	80	70

*Generated by 96 iterations

TABLE II. VALIDATION PERFORMANCE OF DIFFERENT ITERATION NUMBERS WITH BATCH SIZE = 10 AND 20 ON KTH DATABASE

Iterations	**Accuracy (%)**	
	Batch size	
	10	20
950	81.48	84.72
1900	85.19	88.89
3800	86.57	87.96
7600	85.19	89.35
11400	-	89.35

IV. CONCLUSION

In this paper, the feasibility of employing pre-trained AlexNet on human action recognition is explored. The video data is processed through the processes of MHI, MEI and optical flow to compute three different templates that carry spatio-temporal information, motion region information and motion speed information. Then, these templates are inputted into the pre-trained AlexNet for transfer learning by fine-tuning the last three layers. The proposed approach is validated by using Weizmann database. The obtained results show an encouraging performance of our human action recognition system.

ACKNOWLEDGMENT

Authors wish to express their appreciation to Mini Fund Multimedia University for the support.

REFERENCES

[1] H.A. Abdul-Azim and E.E. Hemayed, “Human action recognition using trajectory-based representation,” *Egyptian Informatics Journal* (16), pp. 187-198, 2015.

[2] V. Kellokumpu, G. Zhao, and M. Pietikäinen, “Recognition of human actions using texture descriptors,” *Machine Vision and Applications* (*22)*, pp. 767-780, 2011.

[3] Y. Kong, Y. Fu, “Human action recognition and prediction: A Survey,” arXiv preprint arXiv:1806.11230, 2018.

[4] A. Klaser, M. Marszalek, C. Schmid, “A spatio-temporal descriptor based on 3D-gradient,” in BMVC 2008-19th British Machine Vision Conference, 2008.

[5] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, "Learning realistic human actions from movies", *Proc. Conf. Comput. Vis. Pattern Recogn.*, pp. 1-8, 2008.

[6] Y.H. Pang & E.Y. Khor & S.Y. Ooi, “Localized Temporal Representation in Human Action Recognition,” International Conference on Network, Communication and Computing, pp.261-266, 2018.

[7] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, andL. Fei-Fei, “Large-scale video classification with convolutional neuralnetworks,” inCVPR, 2014.

[8] K. Simonyan, and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” In *Advances in neural information processing systems*, pp. 568-576, 2014.

[9] J. Arunnehru, G. Chamundeeswari, & S.P. Bharathi, “Human Action Recognition using 3D Convolutional Neural Networks with 3D Motion Cuboids in Surveillance Videos,” *Procedia computer science*, *133*, pp. 471-477, 2018.

[10] M. Hasan and A. K. Roy-Chowdhury, “Continuous learning of human activity models using deep nets,” inECCV, 2014.

[11] H. Azizpour, S.A. Razavian, J. Sullivan, "From generic to specific deep representations for visual recognition", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015.

[12] Ghazi, Mostafa Mehdipour, Berrin Yanikoglu, and Erchan Aptoula. "Plant identification using deep neural networks via optimization of transfer learning parameters." *Neurocomputing* 235, pp. 228-235, 2017.

[13] A. Krizhevsky, I. Sutskever, G.E. Hinton, “ImageNet classification with deep convolutional neural networks,” Advances in neural information processing systems (NIPS), pp. 1106–1114, 2012.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, “ImageNet large scale visual recognition challenge”, Int. J. Comput. Vis. 115 (3) (2015), pp. 211–252. http://dx.doi.org/10.1007/s11263-015-0816-y.

[15] A. Abdolmanafi, L. Duong, N. Dahdah, & F. Cheriet, “Deep feature learning for automatic tissue classification of coronary artery using optical coherence tomography,” *Biomedical optics express*, *8*(2), pp. 1203-1220, 2017.

[16] J.W. Davis and A.F. Bobick, “The representation and recognition of human movement using temporal templates,” In *Computer Vision and Pattern Recognition, Proceedings.,* pp. 928-934, 1997.

[17] A.F. Bobick and J.W. Davis, “The recognition of human movement using temporal templates,” *IEEE Transactions on pattern analysis and machine intelligence* (23), pp. 257-267, 2001.

[18] J.L. Barron, D.J. Fleet, S.S. Beauchemin, “Performance of optical flow techniques,” Int. J. Comput. Vis.12, pp. 43–77, 1994.

[19] S. Baker, D. Scharstein, J.P. Lewis, S. Roth, M.J. Black, R. Szeliski,R, “A database and evaluation methodology for optical flow,” Int. J.Comput. Vis.92, pp. 1–31, 2011.

[20] B.D. Lucas, T. Kanade, “An iterative image registration technique with an application to stereo vision,” Proceedings of the 7th International Joint Conference on Artificial Intelligence, pp. 674-679, 1981.