Multiple classifier-based spatiotemporal features for living activity prediction

Van-Dung Hoang

Faculty of Engineering - Technology, Quang Binh University, Dong Hoi, Quang Binh, Vietnam

ABSTRACT
Nowadays, the action prediction technique plays an important role in many automatic systems. There are some proposed methods for this issue. However, they retain limitations such as accuracy and computational time, especially for applying in limited resource systems. This paper presents an approach to enhance the efficiency of the activity prediction task. The work processes on multiple classifiers using spatiotemporal features based on scalable feature descriptors, such as histogram of oriented gradients (HOG), histogram of oriented optical flow (HOF), and motion boundary histogram (MBH). In order to improve prediction accuracy, two layers of classified machine models are studied for applying on spatiotemporal features with the dynamic foreground extraction process. The first layer based on unsupervised classification is proposed to construct a dictionary of features, which supports for distinguishing and uniform the number of features. The next task is that supervised machine learning supports for final decision of action classes. The proposed approach was evaluated on several benchmark datasets, which are available online. The results demonstrate that the approach enhances accuracy and efficiency of the prediction system.

1. Introduction
Nowadays, human detection and action prediction systems based on vision sensors have been considered key tasks in a variety of modern system applications, which is a potential impact on many intelligence and surveillance systems, autonomous systems. They are related to image retrieval, automatic personal assistance, and intelligent transportation. The action recognition task aims to identify some kinds of action class of humans or animals given with contextual conditions. Generally, just one or several kinds of daily living actions are focused on to recognize from the observed series. However, there are still many challenges in action prediction procedures such as various activities, diversity of the field of view, apparent differences, complex backgrounds, as well as occlusion. Even though, not only artificial intelligent systems but also human are confused when to predict action classes, in sometimes. There are usually several tasks for detection and prediction of an action. Two major important tasks should be solved. Object/human is...
detected from the region of interest by analysis of the image texture and structure. Then moving components are extracted to investigate for action prediction.

In order to enhance efficiency in computational cost, we present a new approach on scalable features in pyramid images. The scalable features are described based on histograms’ orientation of gradient and optical flow of the moving component. In this processing, the basic gradient and optical flow are stored in special structures, which supports the extraction of the pyramid features without recomputing that information. To improve accuracy, supervised and unsupervised machine classified models are applied. First, $k$-mean-based unsupervised classification is proposed to construct a dictionary of features, which supports the distinguishing of and uniform the number of features. Second, the Support Vector Machine (SVM)-based supervised machine learning is applied to predict living activities. The general flowchart of action prediction is briefly summarized in Figure 1.

The rest of the paper is organized as follows. Related work is presented in Section 2. Section 3 presents a method for extracting spatiotemporal feature descriptors. Dynamic foreground extraction task supports to extract the region of interesting (ROI) which caused of activities is briefly described in Section 4. Multiple classifiers based on unsupervised and supervised machine learning are presented in Sections 5 and 6. Some evaluation results of the experiment are explained and illustrated in Section 7. The last section presents the discussion and conclusion.

2. Related work

In recent years, some contributions in the field of action recognition have been proposed, such as in Liu, Xu, Qiu, Qing, and Tao (2016), Shi, Laganière, and Petriu (2016), and Stefic and Patras (2016). A summary report of the state-of-the-arts was discussed and presented in González et al. (2015) and Ziaeefard and Bergevin (2015). The literatures’ brief evaluated some advanced methodologies to be applied to action recognition and inferred in experimental results. Some advantages and limitations of them were also pointed out. Besides

![Figure 1. Overview diagram of action recognition processing.](image)
that, some standard datasets are available online as benchmark data for competition and testing, which were analysed and evaluated in Chaquet, Carmona, and Fernández-Caballero (2013), such as KTH (Schuldt, Laptev, & Caputo, 2004), HOLLYWOOD (Marszalek, Laptev, & Schmid, 2009), and UCF101 (Soomro, Zamir, & Shah, 2012). Most of the datasets are collected as RGB information, which is a rich texture for feature extraction. In the present days, the local feature descriptors are attracting many researchers (González et al., 2015; Seo, Kim, De Neve, & Ro, 2016; Wang, Li, & Fang, 2016; Ziaee fard & Bergévin, 2015). Wang et al. (2016) presented a methodology to transform the gradient-based features into spatiotemporal concatenation with Gaussian convolution for extracting spatial location features with the expectation of improving the precision rate. Seo et al. (2016) proposed an integrated technique to reject superfluous trajectories that resulted in camera motion, while still maintaining the effectiveness action prediction results. Some other groups focus on global feature descriptors for improving the result, such as Azis, Jeong, Choi, and Iraqi (2016) and Vishwakarma and Kapoor (2015). Azis et al. (2016) proposed a method based on multi-view for action recognition using a weighted averaging fusion to merge skeletal data from multiple views. Vishwakarma and Kapoor (2015) investigated a method hybrid classification model using SVM and k-Nearest Neighbour (k-NN) using human silhouette and grids for modelling feature vectors. A graph model-based method in Yi and Lin (2016) was proposed for multiple instance learning. Authors used a graph for presenting the local information, which expected faster than the previous subspace learning methods with computational complexity.

In contrast, some research groups focus on RGBD data for action recognition (Jia & Fu, 2016; Kong & Fu, 2016; Zhang, Li, Ogunbona, Wang, & Tang, 2016). They presented methods using the low rank of tensor to automatically learn the subspace dimension. Action recognition in RGBD video (Kong & Fu, 2016) was investigated for improving the accuracy quality based on a discriminative relation by discarding homogeneous pieces in RGB image and utilizing depth information for classifying a set of actions. Some advantages and limitations of action recognition techniques using RGBD images were analysed and presented in Zhang et al. (2016). Approaches based on a combination of colour and deep information were investigated to result in acceptable accuracy. However, the problem is expensive computational cost.

Different to almost above contributions, instead of an individual set of period frames is used for prediction, the part overlapping sequence of period frames are considered to convolute and infer action classes. Modelling object information is very important in order to distinguish action meaning from scenario and context. In this situation, identifying the relative object with the major object is very important to recognize its class. The trajectory of components and geometry relationship are important signals for distinguishing how it is different activities, for example, difference between running, jogging, and walking. In this contribution, we improve the action recognition result by using the spatiotemporal local feature with multiple classification layers for prediction.

3. Spatiotemporal feature extraction

In this study, some kinds of local feature descriptors are studied for applying in pattern recognition, such as HOG, HOF, and MBH. They support classification and recognition,
for example, they are robustness to geometry properties, complex background, independent machine learning and tractable filter. However, there remain some limitations in computational cost and accuracy. We proposed a method for computing the feature descriptors while maintaining the efficiency rate. First, it is necessary to determine feature locations, which consist of rich structure for robust tracking. Local feature descriptors are extracted at every detected location in each image frame. Then, each feature location is tracked within an interval of the action cycle. At end of each action cycle, the spatiotemporal feature is constructed within $L$ consecutive frames. They are fed to the next step of action prediction processing.

In this study, a dense sample feature extraction is generated by cell splitting in multiple scale levels. The input image is divided into the cells with a cell size ($c_w$, $c_h$), where $c_w$, $c_h$ are width and height of the cell, respectively. Point of interest (POI) for extracting features called feature location. It is suggested to discard homogeneous regions with poor structure of content in order to extract robust feature locations for tracking them in sequent frames. Dense optical flows are extracted using the Farneback algorithm in Farnebäck (2003). In that method, the optical flow of two sequent images is computed based on polynomial expansion. In this application, a camera is not too high-frequency vibrations, rotation. Objects smoothly move and the displacement field is slow variation. Therefore, it is not necessary to integrate information over the neighbourhood of each point. It is utilized for reducing computational time. Spatiotemporal feature descriptors are processed over $\sigma_t$ sequent frames, as depicted in Figure 2. The set of points within $L$ frames period is divided into $n$ intervals of $\sigma_t$ frames for extracting and constructing feature descriptors. Feature descriptors within $n$ subsequent intervals are represented as the concatenation of the set spatiotemporal features, where the number $\sigma_t$ is defined as a divisor of $L$ frames. The volume feature is defined as $\sigma_x \times \sigma_y \times (L/\sigma_t)$ elements. For the basic techniques to extract feature descriptors of HOG, HOF, and HMB refer to Dalal and Triggs (2005) and Dalal, Triggs, and Schmid (2006) for more details.

![Figure 2. Spatiotemporal features extraction.](image-url)
4. Dynamic foreground extraction

Motion components of human are important signals, which express of activities. Meanwhile, static components are usually ambiguous for action prediction. Therefore, it is important to extract moving parts, related to the human body. In the task of obtaining foreground from sequent images from a dynamic camera, it is not easy to segment only the moving region of the object of interest, because it faces two kinds of motion, the first is independent motion, which is caused by the object’s movement and the second is motion from the camera motion. The problem of camera motion is solved by the use of optical flow properties in order to segment the independent motion of the moving object based on ego-motion compensation technique (Hariyono, Hoang, & Jo, 2014).

In order to extract dynamic foreground, a background subtraction (BS) technique is applied to the ROI. This process is indicated whether each piece of the foreground is motion or static. Background modelling consists of two main steps, background initialization and background update for adapting to the scene changing over time. Suppose that a set of consecutive ROIs (foreground ROI) is \( r = \{ r_1, r_2, \ldots, r_n \} \), with ROI size \( m \times n \). The first step, initial background model \( B(x, y) \) for each piece, \( 1 \leq x \leq m \) and \( 1 \leq y \leq n \). The second step, model is updated for new frames \( r_t \), classify the pixel \( (x, y) \) into either motion piece \( M_t(x, y) \) or static piece \( S_t(x, y) \). If the piece is classified as static, the static piece \( S(x, y) \) should be updated, which is called background model \( B_m \). These processes are typically conducted in many approaches, Mixture of Gaussian Modelling (MOG) (Amazon) and enhanced Gaussian Mixture Model (EGMM) (FPT). The moving component is mathematically defined as follows

\[
M_t = \begin{cases} 
1 & \text{if } |B_m - r_t| > \text{threshold} \\
0 & \text{otherwise}
\end{cases}
\]

In order to solve the problem of changed illumination and other variety conditions, the model can be iteratively updated in the following form

\[
B_m = \lambda B_m + (1 - \lambda) r_t.
\]

For updating the model, a learning rate \( \lambda \in [0, 1] \) is usually used which gives a trade-off between \( \lambda B \) and \( (1 - \lambda) r_t \). For a smaller value of \( \lambda \), the model changes faster, while for a bigger value of \( \lambda \), the model is retained longer (Figure 3).

5. Bag of words-based classification

A classifier based on the framework of BoW is discussed and proposed for application in order to distinguish the feature property and reduce the dimension of the feature descriptor. Usually, local descriptors are high dimensional and strong correlated. This is a great challenge in processing time and accuracy of the systems. There are several steps to construct the BoW model. Spatiotemporal feature descriptors are extracted from training video data. The set of the feature vectors is fed to train the BoW model. In this experiment, unsupervised learning is proposed to do this task. Due to a variety of activity situations and environment conditions, the number of features extracted at every frame of action cycle is different. Classification task supports to unify domain of feature vector to feed to next
prediction step. Additionally, constructing distinguished features with low dimension is an essential criterion to reduce the computational complexity and improve the accuracy of the prediction system.

In classification and quantization of data to construct the BoW model, there are many applicable unsupervised learning methods such as $k$-means, density-based spatial clustering of applications with noise (DBSCAN), hierarchical Gaussian mixture models (GMM), and so on. The extended $k$-means algorithm is recommended for application. This is because it is simple and fast for use in this task and produces competitive results.

Given a set of high dimension feature vectors $\{x_i | x_i \in \mathbb{R}^n, i = 1 \ldots m\}$, $k$ words of the BoW dictionary. Each word as a clustering represents each group of strong correlation features. The dimension number $n$ depends on the kind of descriptors. For example, 96 elements concerns HOG features, which formed by 32 elements for each POI. Spatiotemporal feature are extracted in a period of 5 frames with consists 3 intervals of the spatio-temporal descriptor. The set of training data is used to generate $k$ clustering of BoW. That means a BoW dictionary consist of $k$ words. The objective is to minimize a potential function with an applicable computational time of $k$-words. The greedy method is one of the most popular methods to reach minimized functional cost. However, it encounters the problem of computing complexity. Due to huge data size and high dimension, the method in Arthur and Vassilvitskii (2007) is used to generate initial words of the dictionary. They are chosen in a specific way to easily gain the error criterion.

$$\arg\min_{D} \sum_{j=1}^{k} \sum_{x \in c_j} ||x - w_j||^2,$$

where $\{w_i | i = 1, \ldots, k\}$ is a word in the BoW dictionary, $X = \{x_i | x_i \in \mathbb{R}^n\}$ is a set of elements that belong to the $j$th clustering $c_j$ that mean all vectors $x_i$ vote for their clustering centre $w_j$, $D$ is a domain of dictionary.

In BoW feature extraction task, there are many approaches for matching and computing the BoW feature, such as linear searching, greedy method, Brute-Force matching, RANSAC,
and so on. The amount of methods, a matching algorithm of fast approximate nearest neighbours (FANN, Muja & Lowe, 2009) is a significant solution. In that approach, classification is processed by use of some different algorithms for approximate nearest neighbour search on the set of big data with high dimension. The experimental result in Muja and Lowe (2009) also indicated that it archives high accuracy, which is evaluated on the percentage of points for the correct nearest neighbour based on the hierarchical k-means tree and multiple randomized kd-trees. The objective is the optimization problem of finding the closest word in the dictionary, which is formed as follows

$$\text{argmin}_{w_j \in X} \sum_i ||x - w_j||^2,$$  \hspace{1cm} (2)

where \(x\) is a input feature vector in \(n\) dimension, \(w_j\) is the closest centre to the sample \(x\).

The BoW-based feature \(B(b_1, b_2, \ldots, b_k)\) is a voting of all samples to the BoW dictionary. Finally, the extracted feature vector is normalized in the following form

$$f = \left[ \frac{b_1, b_2, \ldots, b_k}{\sum_{i=1}^{k} b_i} \right].$$ \hspace{1cm} (3)

6. Supervised learning for detection

This section presents an approach based on a supervised machine approach to handle the problem of action prediction. To deal with this problem, this paper uses an SVM technique for training action models. The details of SVM could be referred to Chih-Chung and Chih-Jen (2011), it becomes a state of the art technique and successfully implemented. SVM has been proven and applied successfully in many fields. The main advantage of the SVM technique is its ability to extract highly discriminative profits. This contribution focuses on this problem based on associated component detections and action prediction. Given the training set, which consists of \(D = \{(u_i, y_i) | i = 1 \ldots n\}\), \(u_i\) is a feature vector of the ROI sample component or negative sample, and label \(y_i\) indicates action class such that \(y_i \in \{\text{domain of actions}\}\). In order to apply to multiple classes of actions, the multiple binary classifications are processed in the cascade structure. The objective SVM training solves the primal optimization hyperplane for maximum margin classification to each class, which is expressed as follows

$$\min_{w, b, \xi} \frac{1}{2} ||w||^2 + C \sum_{i=1}^{n} \xi_i,$$ \hspace{1cm} (4)

s.t. \(y_i(w^T \phi(u_i) + b) \geq 1 - \xi_i \quad \xi_i \geq 0, \quad i = 1, \ldots, n\)

where \(\phi(x_i)\) is a mapping vector \(x_i\) into a higher-dimensional space for linear classification, and \(C > 0\) is a regularization penalty parameter assigning to error coefficient with the optimal \(w\) satisfying

$$w = \sum_{i=1}^{ns} y_i \alpha_i \phi(u_i).$$

After the training step, parameters of the classification model are stored, for example, coefficient \(b\), the set of support vectors with respect to their label \((u_i, y_i), i = 1, \ldots, ns\) as well as positive Lagrange multiplier coefficients \(\{\alpha_i | \alpha_i \geq 0, \text{ with } i = 1 \ldots ns\}\) with \(ns\) being the number of support vectors. This experiment uses the Gaussian radial basis function (RBF) kernel, which is defined as \(k(u_i, u) = \exp (-||u - v||^2 / \sigma^2)\). The signed distance of
feature vector \( \mathbf{u} \) to the hyperplane margin of the SVM model is presented as follows

\[
h(\mathbf{u}) = \sum_{i=1}^{n_s} y_i \alpha_i k(\phi(\mathbf{u}), \phi(\mathbf{u})) + b.
\]  

(5)

The output probability of the action prediction is formulated as

\[
\Phi(\mathbf{u}) = P(y = 1 | \mathbf{u}) = \frac{1}{1 + \exp(-h(\mathbf{u}))},
\]

where \( \mathbf{u} \) is a feature vector which represents the action class.

7. Experiment

We evaluated the proposed approach on two benchmark datasets. First, the KTH dataset (Schuldt et al., 2004) consists of six action classes such as Boxing, Handclapping, Hand waving, Jogging, Running, and Walking. They were obtained by recording 25 people for 6 classes of activities. The dataset was collected under four different scenarios, such as static and homogenous background (d1), scale variations (d2), apparent differences (d3), and lighting variations (d4). This dataset includes 100 videos for each class. Totally, there are 600 videos of 6 action classes, with each recorded situation including 25 videos/class. The configuration of the video is with 25 fps and 160 × 120 pixels resolution. Each person did frequent exercise of only one action class in a video. Figure 4 shows some videos, which were used for the experimental evaluation. The actions of the KTH dataset are quite uniform appearance of human, unrealistic activities (acting), background and selected scenarios and environment.

Spatiotemporal features are extracted in an interval of 15 frames for predicting of an action class. There are two layers of classification. The first classifier is based on BoW for distinguishing and uniform the number of features of each action. The results are fed to SVM for final prediction. Table 1 illustrates the prediction results on the KTH dataset.
Some activities are similar, which caused missed prediction such as running, jogging, and walking. The motion trajectories of legs, arms, and body are similar. However, the relative geometry between two legs of person is an important signal for distinguishing by the first classification layer. Figure 5 shows some prediction results on the KTH dataset.

The second experiment was processed on a dataset from the University of Central Florida (Soomro et al., 2012), called UCF101, which consists of 101 action classes. The UCF101 dataset includes realistic action videos, which were retrieved from YouTube. Totally, there are 13,320 videos of the total 101 action categories. The set of videos is with a diversity object appearance and pose, object scales, viewpoint, cluttered background, illumination conditions, camera motion, and so on. Videos were unified with the frame rate of 25 fps and video resolution of 320 × 240 pixels.

We categorized the following topics into three subsets: instrument playing action (12 action classes), daily living action (20 action classes), and another action (69 classes). The first dataset relates to instrument playing activities, which consists of ‘Band Marching’, ‘Drumming’, ‘Playing Cello’, ‘Playing Daf’, ‘Playing Dhol’, ‘Playing Flute’, ‘Playing Guitar’, ‘Playing Piano’, ‘Playing Sitar’, ‘Playing Tabla’, and ‘Playing Violin’. The second dataset relates to daily living activities consisting of ‘Apply Eye Makeup’, ‘Apply Lipstick’, ‘Baby Crawling’, ‘Biking’, ‘Blow Dry Hair’, ‘Blowing Candles’, ‘Brushing Teeth’, ‘Cutting In Kitchen’, ‘Haircut’, ‘Hammering’, ‘Head Massage’, ‘Knitting’, ‘Military Parade’, ‘Mixing’, ‘Mopping Floor’, ‘Shaving Beard’, ‘Trampoline Jumping’, ‘Typing’, ‘Walking With Dog’.

Table 1. Confusion matrix of prediction results on the KTH dataset.

|        | Boxing | Clapping | Waving | Jogging | Running | Walking |
|--------|--------|----------|--------|---------|---------|---------|
| Boxing | 0.9899 | 0.0050   | 0.0012 | 0.0003  | 0.0001  | 0.0035  |
| Clapping | 0.0015 | 0.9949   | 0.0036 | 0         | 0       | 0       |
| Waving  | 0.0040 | 0.0195   | 0.9765 | 0.8571   | 0.1090  | 0.0317  |
| Jogging | 0.0016 | 0.0003   | 0.0003 | 0.1892   | 0.7983  | 0.0122  |
| Running | 0.0001 | 0.0002   | 0.0001 | 0.0328   | 0.0199  | 0.9457  |

Figure 5. Some results of action prediction on the KTH dataset.
Figure 6. Crosschecked matrix of prediction results on some datasets.
and ‘Writing On Board’. The third dataset about sport and other activities consists of 69 action classes. They are labelled as ‘Baseball Pitch’, ‘Basketball’, ‘Bench Press’, ‘Biking’, ‘Billiards’, ‘Breast Stroke’, ‘Clean And Jerk’, ‘Diving’, ‘Drumming’, ‘Fencing’, ‘Golf Swing’, ‘High Jump’, ‘Horse Race’, ‘Horse Riding’, ‘Hula Hoop’, ‘Javelin Throw’, ‘Juggling Balls’, ‘Jump Rope’, ‘Jumping Jack’, ‘Kayaking’, ‘Lunges’, ‘Military Parade’, ‘Mixing’, ‘Nunchucks’, ‘Pizza Tossing’, ‘Playing Guitar’, ‘Playing Piano’, ‘PlayingTabla’, ‘Playing Violin’, ‘Pole Vault’, ‘Pommel Horse’, ‘Pull Ups’, ‘Punch’, ‘Push Ups’, ‘Rock Climbing Indoor’, ‘Rope Climbing’, ‘Rowing’, ‘Salsa Spin’, ‘Skate Boarding’, ‘Skiing’, ‘Skijet’, ‘Soccer Juggling’, ‘Swing’, ‘TaiChi’, ‘Tennis Swing’, ‘Throw Discus’, ‘Trampoline Jumping’, ‘Volleyball Spiking’, ‘Walking With Dog’, and ‘YoYo’. Videos are realistic actions, which were collected from films on YouTube.

The confusion matrix of predicted results is shown in Figure 6. Experimental results illustrated that efficiency of the prediction system reached a high accuracy about 95.3964% on UCF12, about 84.4232% on UCF20 with the use of 2000 words of BoW, and 1000 samples for SVM training. On UCF69, the system archives about 78.1261% accuracy rate with parameter setting of 4000 words of BoW and 1000 samples for training.

We also evaluated on the different size of the BoW dictionary and the number of samples used for SVM model training. The experimental results are shown in Table 2. Actually, the results illustrated that the bigger BoW dictionary and more samples for SVM training archived higher performance rate and vice versa. However, there is a trade-off between accuracy and computational time on BoW size and number of samples for training.

### Table 2. Comparison results of action prediction using different size of BoW and the number of samples for model training.

| Dataset | #BoW dictionary | #Samples for SVM | Accuracy rate (%) |
|---------|-----------------|-----------------|------------------|
| UCF12   | 2000            | 1000            | 95.3964          |
|         | 2000            | 2000            | 96.6845          |
|         | 2000            | 3000            | 97.2462          |
| UCF20   | 2000            | 1000            | 84.4232          |
|         | 2000            | 2000            | 88.3467          |
|         | 2000            | 3000            | 89.8813          |
| UCF69   | 2000            | 1000            | 76.5808          |
|         | 2000            | 2000            | 81.1492          |
|         | 2000            | 3000            | 81.6430          |
|         | 4000            | 1000            | 78.1261          |
|         | 4000            | 2000            | 83.1576          |
|         | 4000            | 3000            | 83.9431          |

8. Conclusion

This paper presents a framework for action prediction based on multiple classifiers using spatiotemporal local features. There are several tasks of the system. First, local features are extracted at every frame from consecutive images. Second, the spatiotemporal feature descriptors are constructed based on local features in each interval of $L$ frames. Third, a set of spatiotemporal feature vectors is classified for distinction and uniform the number of features of each candidate action. The final action class is predicted by using linear SVM. The prediction approach was successfully implemented for testing on some benchmark datasets, such as KTH and UCF101. Experiment results proved that the
investigated method outperforms. Future work will focus on reducing computational time for local feature extraction with expected to real-time applications in the limited resource systems.

**Disclosure statement**

No potential conflict of interest was reported by the author.

**Funding**

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.05-2015.09.

**Notes on contributor**

*Van-Dung Hoang* received his Ph.D. degree from University of Ulsan, Korea, in 2015. Since 2002, he has been serving as a lecturer in University of Quang Binh, Vietnam. He is actively participating as a member of the societies as IEEE, ICROS. His research interests include pattern recognition, machine learning, computer vision, databases, cryptography.

**ORCID**

*Van-Dung Hoang* [http://orcid.org/0000-0001-7554-1707](http://orcid.org/0000-0001-7554-1707)

**References**

Amazon. *Thiết bị và cờ cụ hỗ trợ lập trình alexa ứng dụng Smart-home.* [https://developer.amazon.com/alexa.](https://developer.amazon.com/alexa.)

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 1027–1035). New Orleans, Louisiana.

Azis, N. A., Jeong, Y. S., Choi, H. J., & Iraqi, Y. (2016). Weighted averaging fusion for multi-view skeletal data and its application in action recognition. *IET Computer Vision, 10*(2), 134–142.

Chaquet, J. M., Carmona, E. J., & Fernández-Caballero, A. (2013). A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding, 117*(6), 633–659.

Chih-Chung, C., & Chih-Jen, L. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*(3), 1–27.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 886–893). Retrieved from [http://ieeexplore.ieee.org/abstract/document/1467360/?reload=true](http://ieeexplore.ieee.org/abstract/document/1467360/?reload=true)

Dalal, N., Triggs, B., & Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *9th European Conference on Computer Vision* (pp. 428–441). Retrieved from [http://link.springer.com/chapter/10.1007/11744047_33](http://link.springer.com/chapter/10.1007/11744047_33)

Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis* (pp. 363–370).

FPT. *Tập đoàn FPT Telecom, Camera đảm bảo an toàn giao thông.* [http://fpt.center/.](http://fpt.center/)

González, S., Sedano, J., Villar, J. R., Corchado, E., Herrera, Á, & Baruque, B. (2015). Features and models for human activity recognition. *Neurocomputing, 167*, 52–60.

Harlyono, J., Hoang, V.-D., & Jo, K.-H. (2014). Motion segmentation using optical flow for pedestrian detection from moving vehicle. In *International Conference on Computational Collective Intelligence* (pp. 204–213). Retrieved from [http://link.springer.com/chapter/10.1007/978-3-319-11289-3_21](http://link.springer.com/chapter/10.1007/978-3-319-11289-3_21)
Jia, C., & Fu, Y. (2016). Low-rank tensor subspace learning for RGB-D action recognition. *IEEE Transactions on Image Processing, 25*(10), 4641–4652.

Kong, Y., & Fu, Y. (2016). Discriminative relational representation learning for RGB-D action recognition. *IEEE Transactions on Image Processing, 25*(6), 2856–2865.

Liu, F., Xu, X., Qiu, S., Qing, C., & Tao, D. (2016). Simple to complex transfer learning for action recognition. *IEEE Transactions on Image Processing, 25*(2), 949–960.

Marszalek, M., Laptev, I., & Schmid, C. (2009). Actions in context. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2929–2936). Retrieved from http://ieeexplore.ieee.org/document/5206557/

Muja, M., & Lowe, D. G. (2009). Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1), 2*(331–340), 2.

Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: A local SVM approach. In *The 17th International Conference on Pattern Recognition* (pp. 32–36). Retrieved from http://ieeexplore.ieee.org/abstract/document/1334462/

Seo, J.-J., Kim, H.-I., De Neve, W., & Ro, Y. M. (2016). Effective and efficient human action recognition using dynamic frame skipping and trajectory rejection. *Image and Vision Computing, 58*, 76–85. doi:10.1016/j.imavis.2016.06.002.

Shi, F., Laganière, R., & Petriu, E. (2016). Local part model for action recognition. *Image and Vision Computing, 46*, 18–28.

Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. *CRCV-TR-12-01.*

Stefic, D., & Patras, I. (2016). Action recognition using saliency learned from recorded human gaze. *Image and Vision Computing, 52*, 195–205.

Vishwakarma, D. K., & Kapoor, R. (2015). Hybrid classifier based human activity recognition using the silhouette and cells. *Expert Systems with Applications, 42*(20), 6957–6965.

Wang, L., Li, R., & Fang, Y. (2016). Gradient-layer feature transform for action detection and recognition. *Journal of Visual Communication and Image Representation, 40*(Part A), 159–167.

Yi, Y., & Lin, M. (2016). Human action recognition with graph-based multiple-instance learning. *Pattern Recognition, 53*, 148–162.

Zhang, J., Li, W., Ogunbona, P. O., Wang, P., & Tang, C. (2016). RGB-D-based action recognition datasets: A survey. *Pattern Recognition, 60*, 86–105.

Ziaeefard, M., & Bergevin, R. (2015). Semantic human activity recognition: A literature review. *Pattern Recognition, 48*(8), 2329–2345.