Sensors & Transducers, Vol. 167, Issue 3, March 2014, pp. 23-29

Sensors & Transducers

Human Action Recognition Based on Boosting

Zhu Shao-Ping
Department of Information Management, Hunan University of Finance and Economics,
Changsha, 410205, China
Tel.: 13787247056
E-mail: zhushaoping_cz@163.com

Received: 20 December 2013 /Accepted: 28 February 2014 /Published: 31 March 2014

Abstract: Human action recognition is an active research field in computer vision and image processing. In this paper we propose a novel method for the task of recognition of human actions in video image sequences. First of all, a video sequence is represented as a collection of spatial-temporal words by extracting space-time interest points, which is used to characterize action. Then visual words are used to represent human actions by using Bag of Words. Final boosting algorithm is used for human actions recognition. We test our algorithm on a challenging dataset: the KTH human action dataset. The experimental results show that the average recognition accuracy is over 88 %, which validates its effectiveness. *Copyright © 2014 IFSA Publishing, S. L.*

Keywords: Human action, Boosting algorithm, Space-time interest points, Bag of Words.

1. Introduction

Recent advances in computer vision and pattern recognition have fuelled numerous initiatives to intelligently recognize human actions. Recognizing of human actions in video sequences is essential for various applications, including video surveillance system, video indexing system and human computer interactions. Specifically, a human action recognition system may enable the detection of abnormal actions as opposed to the normal action of persons, which uses at public places like airports, supermarkets and subway stations. Automated human action recognition may be useful for real-time monitoring of the elderly people, patients, or babies. In particular, human action recognition aims at automatically telling the activity of a person, i.e. to identify if someone is walking, running, jogging, dancing, or performing other types of actions. It is a crucial prerequisite for a number of applications, like video surveillance, object-level video summarization, content based image retrieval, digital library organization, etc. However, it remains a challenging problem in computer vision that we achieve robust human actions recognition from image sequences to due to occlusion, articulation in poses, changing backgrounds, camera movements, photometric and geometric variances of objects. With moving target, moving cameras, and non-stationary background, few vision algorithms could categorize and recognize human actions well.

In this paper, we propose a method for automatically recognizing human action from video sequences. This approach includes two steps: extracting feature of human action and recognizing human action. In the extracting feature, features of human action are extracted by extracting space-time interest points. Then we convert features of human actions to visual words using "bag-of-words" (BOW) models [1, 2], and human action is represented by a number of visual words.

Final boosting algorithm is used for human action recognition.

The rest of this paper is organized as follows. Section 2 gives a brief survey of some recent work on human action recognition. After reviewing related work, we describe the feature of human action extraction base on spatial-temporal words and "bag-of-words" models in Section 3. Section 4 gives details of boosting algorithm for recognize human action. Section 5 shows experiment result, also comparing our approach with two state-of-the-art methods, and the conclusions are given in the final section.

2. Related Work

A lot of previous work has been presented to recognize human actions. Specifically, Bayesian approaches and Hidden Markov Models (HMM) have been extensively used to detect simple and complex events that occur in the scenarios. e. g., Olivera *et al.* [3], Bobick and Wilson [4], Xiang and Gong [5], they try to model the full dynamics of videos using sophisticated probabilistic models. The problem with this approach is that those sophisticated models impose too many assumptions and constraints (e. g., the independence assumption of hidden Markov models) in order to be tractable. It is also hard to learn those models since there are usually a large number of parameters that need to be set. One popular approach is to apply tracked motion trajectories of body parts to action recognition [6, 7]. This is done with much human supervision and the robustness of the algorithm is highly dependent on the tracking system. Bobick and Davis [8] use a representation known as "temporal templates" to capture both motion and shape, represented as evolving silhouettes. P. Flaherty *et al.* [9] propose a LDA method of motion processing for action recognition. Christian Schuldt *et al.* [10] perform human action recognition by training SVM classifiers. But these approaches ignore the contextual information provided by different frames in a video the modeling and learning frameworks are rather simple. Another work named video epitomes is proposed by Cheung *et al.* [11]. They model the space-time cubes from a specific video by a generative model. The learned model is a compact representation of the original video, therefore this approach is suitable for video super-resolution and video interpolation, but not for recognition.

A lot of work has been done in recognizing actions from both still images and video sequences. In this paper, we focus on recognition based on "bag-of-words" models and boosting algorithm. We use spatial-temporal words to obtain from the whole frame to create the "visual words." We will demonstrate that "bag-of-words" models and boosting algorithm are better suited for the task of recognizing human actions.

3. Feature Representation from Space-Time Interest Points

We represent each video sequence as a collection of spatial temporal words by extracting space-time interest points. There is a variety of methods for interest points detection in image [12]. Laptev and Lindeberg [13] proposed an extended version of the interest points detection in the spatial domain [14] into space-time domain by requiring image values in space-time to have large variations in both dimensions. Blank *et al.* [15] represented actions as space-time shapes and extracted space-time features such as action dynamics, local space-time saliency, shape structures and orientation for action recognition. As noticed in [16] and from our experience, the interest points detected using the generalized space-time interest point detector are too sparse to characterize many complex videos. Therefore, we used HOG and T-HOG method to detect space-time interest points [17, 18] and use BOW method to obtain action feature vector. Here we give a brief review of this method.

The step for characterizing human action is as follows:

Step 1: Given a stabilized video sequences $f(x,y)$, we detect space-time interest points.

$f(x,y,t)$ is the image in pixel (x,y) at time t, $g(x,y,t,\sigma_l^2,\tau_l^2)$ is Gaussian function that time and space parameters can be separated. Linear multi-scale space is defined as:

$$L(x,y,t;\sigma_l^2,\tau_l^2)=\mathrm{g}(x,y,t;\sigma_l^2,\tau_l^2)*f(x,y,t)\,, \quad (1)$$

where * is the convolution operator, σ_l^2,τ_l^2 is the independent space scale variable and independent time scale variable.

We defined Gaussian function as:

$$\mathrm{g}(x,y,t;\sigma_l^2,\tau_l^2)=\frac{1}{\sqrt{(2\pi)^3\sigma_l^4\tau_l^2}}\times\exp\left(-(x^2+y^2)/2\sigma_l^2-t^2/2\tau_l^2\right) \quad (2)$$

Space-time second moment matrix μ is constructed to detect space-time interest points by convolution of multi-scale space L and Gaussian weighting function, which is expressed as:

$$\mu(\cdot;\sigma_l^2,\tau_l^2)=\mathrm{g}(\cdot;\sigma_i^2,\tau_i^2)*\left(\nabla L(\cdot;\sigma_l^2,\tau_l^2)(\nabla L(\cdot;\sigma_l^2,\tau_l^2))^T\right), \quad (3)$$

where $(\cdot)$ is the simplification of (x,y,t).

$$\begin{aligned} L_x\left(\cdot;\sigma_l^2,\tau_l^2\right)&=\partial_x(g*f),\\ L_y\left(\cdot;\sigma_l^2,\tau_l^2\right)&=\partial_y(g*f),\\ L_t\left(\cdot;\sigma_l^2,\tau_l^2\right)&=\partial_t(g*f). \end{aligned} \quad (4)$$

$\sigma_i^2 = s\sigma_l^2, \tau_i^2 = s\tau_l^2$, $\nabla L(\cdot;\sigma_l^2,\tau_l^2)$ are the first derivatives of scale space function L respectively in the x, y, t direction.

The gradient matrix of multi-scale space L in all directions is expressed as:

$$\nabla L(\cdot;\sigma_l^2,\tau_l^2)(\nabla L(\cdot;\sigma_l^2,\tau_l^2))^T = \begin{pmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2 \end{pmatrix} \quad (5)$$

In this paper we use threshold function H to detect space-time interest points. Threshold function H is defined as:

$$H = \det(\mu) - k\,\mathrm{trace}^3(\mu) \quad (6)$$

Assuming $\alpha = \lambda_2/\lambda_1$, and $\beta = \lambda_3/\lambda_1$, λ_1, λ_2, λ_3 are eigenvalues of the second moment matrix μ . Threshold function H is expressed as:

$$H = \lambda_1^3\left(\alpha\beta - k(1+\alpha+\beta)^3\right), \quad (7)$$

where $k \le \alpha\beta/(1+\alpha+\beta)^3$, and a maximum of k is 1/27. The space-time interest point in images $f(x,y,t)$ is the local where H is the local maxima of time and space. Fig. 1 shows examples of the area of space-time interest points for human actions for the KTH dataset.

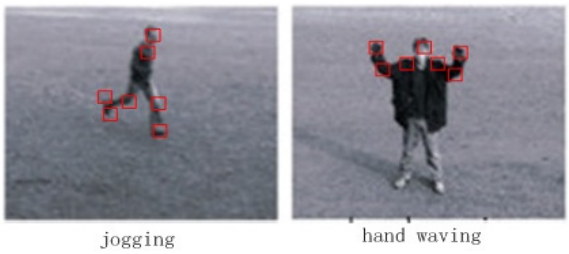

Fig. 1. The area of space-time interest points for human actions.

Step 2: Gradient descriptor HOG and T-HOG are used to describe space-time interest point area cube.

In STIP area cube of each image block frame f_{t+1}, the we calculate size and direction of each pixel gradient as follows:

$$m_{t+i}(x,y) = \sqrt{(L(x+1,y)-L(x-1,y))^2 + (L(x,y+1)-L(x,y-1))^2} \quad (8)$$

$$\theta(x,y) = \arctan\frac{L(x,y+1)-L(x,y-1)}{L(x+1,y)-L(x-1,y)} \quad (9)$$

Step 3: We calculated time direction changes of the cube of space-time interest points. Scale space gradient between frames is defined as:

$$D(x,y,t) = L(x,y,t+1) - L(x,y,t) \quad (10)$$

Size and direction of scale space gradient difference between frames are calculated as follows:

$$q(x,y,t) = \sqrt{D(x,y,t)^2 + D(x,y,t+1)^2} \quad (11)$$

$$\varphi(x,y,t) = \arctan\frac{D(x,y,t+1)}{D(x,y,\mathrm{t})} \quad (12)$$

In this paper, we statistic each frame pixel gradient in the same histogram. In this way, the cube space direction change information can be expressed as a feature vector, which are displayed in Fig. 2.

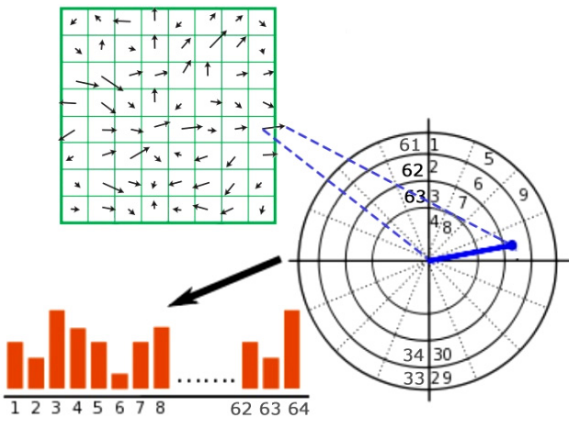

Fig. 2. Gradient-direction descriptor.

Step 4: We construct the codebook randomly by selecting a subset from all STIP area cubes of each image block, then, we use k-means clustering algorithms to obtain V clusters. Codewords are then defined as the centers of the obtained clusters, namely visual words. In the end, each image is converted to the "bag-of-words" representation by appearance times of each codeword in the image is used to represent the image, namely BOW histogram.

Human action is represented by BOW histogram X :

$$X = \{n(I,w_1),\cdots,n(I,w_j),\cdots,n(I,w_M)\}, \quad (13)$$

where $n(I,w_j)$ is the number of visual word w_j included in image, M is the number of vision words in word sets.

4. Human Action Recognition Based on Boosting

4.1. Boosting Fuzzy Classification Algorithm

Boosting fuzzy classification algorithm is a mathematical model based on a fuzzy rule system.

A fuzzy rule system may be defined as follows using the classical case as a beginning. In the classical case, a rule is a function formulated with arguments coupled by logical operators, yielding a logical expression and a corresponding response. If the conditions of the rule are fulfilled (the logical expression is true) then the response has to be true. The logical expression is usually formulated with bivariate logical operators.

With a fuzzy rule, binary logic is replaced by fuzzy logic where a statement and its opposite may both be "true" to different degrees. A fuzzy rule consists of a set of arguments in the form of fuzzy sets with membership functions and a response also in the form of a fuzzy set [19]. Let us consider a set of training input data $X=\{x_1, x_2, \ldots, x_n\}$. For a general input vector, the rule is applied as:

$$R_j: \ \ if\ x_1\ is\ A_{1j}\ and\ x_2\ is\ A_{2j}\ and \cdots and\ x_N\ is\ A_{nj}\ then\ Y = c_j,$$

where x_i is the i-th attribute of the input vector X, and Y is the classification label, which can be considered as an output variable, c_j is the classification, and $c_j \in \{c_1, \cdots, c_m\}$, $j = 1, 2, \cdots, N$. A_{nj} is the fuzzy set of x_j and $\mu_{R_j}(x_j)$ is the subordinate function. We assume that the subordinate function is Gaussian function as follows:

$$\mu_{A_{ij}} = \exp\left[-\frac{1}{2}\left(\frac{x_i - \overline{x}_i^j}{\sigma_i^j}\right)^2\right] \tag{14}$$

For any input variables $\{x_1, x_2, \ldots, x_n\}$, the rules of R_j for excitation degrees is given by

$$\mu_{R_j}(x_j) = \mu_{R_j}(\{x_1, \ldots, x_n\}) = \min_{i=1}^{n} \mu_{A_{ij}}(x_j) \tag{15}$$

The category of R_j can be expressed as:

$$C_{\max}(x_j) = \arg\max_{C_m} \sum_{R_j / c_j = C_m} \mu_{R_j}(x_j) \tag{16}$$

A set of fuzzy classification rules [20, 21] from the given training input data are obtained by parameters of the subordinate function. Before the fuzzy rules are mainly determined by using the neural network method. The neural network method is usually slow, order dependent and incomprehensible. We use Boosting and Genetic algorithms, which generate fuzzy classification rules. Assuming a set of training input data be independent of each other, we propose a genetic algorithm for Boosting fuzzy classification rules determination as follows.

Given a set of training input data $\{(x^1, c_1), (x^2, c_2), \cdots, (x^N, c_N)\}$, and $c_N \in \{c_1, \cdots, c_m\}$, ω^i is the initial weight value calculated by $\omega^i = 1/N$. Genetic algorithm for Boosting fuzzy classification rules determine: $For \quad t = 1, 2, \cdots T \quad Do$. We define as follows:

$$f_1 = \frac{\sum_{k|c^k = c_t} \omega^k \mu_{R_t}(x^k)}{\sum_{k|c^k = c_t} \omega^k}, \tag{17}$$

$$f_2 = \frac{\sum_{k|c^k \neq c_t} \omega^k \mu_{R_t}(x^k)}{\sum_k \omega^k \mu_{R_t}(x^k)}, \tag{18}$$

where ω^k is the k-th weight value of the training sample set. f_1, f_2 are the fitness functions. We calculate

$$f = \begin{cases} 0 & f_2 > k_{\max} \\ f_1 * (1 - \frac{f_2}{k_{\max}}) & f_2 \leq k_{\max} \end{cases}, \tag{19}$$

where $k_{\max}$ is the set as 0.5. We use Genetic algorithm to obtain a fuzzy rule R_t, which is calculated by the maximum f value. When the greater f_1 is and the smaller f_2 is, the greater f is and the smaller classification error rate E_t of fuzzy rules R_t.

Under the current sample distribution, we assume that there are classification error rate E_t of fuzzy rules R_t and the corresponding weights α_t of the rule R_t. We calculate:

$$E(R_t) = \frac{\sum_{i|C_i = C_t} \omega^i \mu_{R_t}(x^i)}{\sum_i \omega^i \mu_{R_t}(x^i)} \tag{20}$$

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1 - E(R_t)}{E(R_{t)}}\right) \tag{21}$$

Assuming the normalized factor z_t, we update sample weights according to error rate as follows:

$$\omega^i(t+1) = \frac{\omega^i(t)}{z_t} \times \begin{cases} e^{-\alpha_t \mu_{rt}(x^i)} & c_i = c_t \\ e^{\alpha_t \mu_{R_t}(x^i)} & c_i \neq c_j \end{cases} \tag{22}$$

For unknown sample $x^k = \{x_1^k, x_2^k, \cdots, x_N^k\}$, the category are obtained by the fuzzy classifier as follows:

$$C_{\max}(x^k) = \arg\max_{C_m} \sum_t \alpha_t \sum_{R_t / c_t = C_k} \mu_{R_t}(x^k) \tag{23}$$

The Genetic algorithm mimics the process of natural evolution, using the survival of the fittest and natural selection principles for tackling classification and optimization problems. They attempt to obtain an optimum result by

swapping parts, selectively mutating chromosomes that encode the solution and evaluate candidate combinations against a fitness function. This procedure has been proved to be effective, as it is used in natural evolution and is extensively used in fuzzy genetic applications. As Boosting method was adopted, every learning fuzzy rules are mainly to the current rule set the training sample that cannot be classified correctly, The new rules have good complementary and be helpful for classification. Reflects the different rules in the classification is different, Boosting method uses the weighted voting classification criterion, the Classification accuracy is higher than the general fuzzy classification (e. g., formulae (16)).

4.2. Human Action Recognition Based on Boosting

We use the Boosting algorithm to learn and recognize human actions. In human actions recognition, it may consist of six human actions (e. g. "walking", "jogging", "running", "boxing", "hand waving" and "hand clapping"). The sample is represented by HOG and T-HOG method to obtain the feature vector and as input of classifier. Assuming the observed data be independent of each other, we use Genetic algorithm for Boosting fuzzy classification and recognize human actions as follows.

INPUTS:

X - feature of human actions

OUTPUTS:

$C_{\max}(X)$ -the category of human actions

Step 1: Extract feature of human actions by gradient descriptor based on HOG and T-HOG and convert actions information to visual words using "bag-of-words" models. Descriptor is used to describe space-time interest point area cube.

Step 2: Using the method of Boosting access to fuzzy classification rule set.

Assuming the training sample set $\{(x^1,c_1),(x^2,c_2),\cdots,(x^N,c_N)\}$, $c_N \in \{c_1,\cdots,c_m\}$. It consists of feature vector X of human action, which categories of human actions are known in a stabilized video sequence in which a person appears in the center of the field of view.

Give equal initial weights of each sample: $\omega^i = 1/N$, the training sample set trains for T rounds of training and obtains T fuzzy classification rules.

For t = 1, 2, …, T Do

Find out a fuzzy rule R_t to maximize the fitness f by using genetic algorithm.

Under the current sample distribution, calculated the corresponding weights of fuzzy rules by the formulae (21).

Update the sample weight by formulae (22).

Step 3: Human actions recognition based on Boosting.

Extract human actions characteristics of the unknown image and get the feature vector X.

Calculate each of the excitation of fuzzy rules R_j by formulae (15).

Determine the category of the human actions by formulae (23).

5. Experimental Results and Analysis

The effectiveness of the proposed algorithm was verified by using C++ and Mat lab hybrid implementation on a PC with Pentium 3.2 GHz processor and 4 G RAM.

We test our algorithm using KTH human motion dataset, which is the largest available video sequence dataset of human actions [22]. In this database, there are six groups of images ("walking", "jogging", "running", "boxing", "hand waving" and "hand clapping") by 25 subjects in different scenarios of outdoor and indoor environment with scale change. Sample images are shown in Fig. 3.

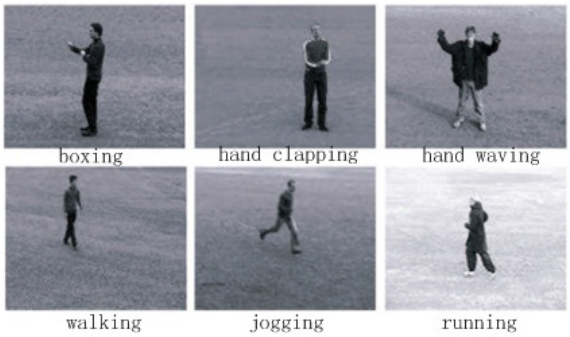

Fig. 3. Key frames for KTH data set.

In this experiment, we study recognition accuracy of six kinds of human actions from KTH human motion dataset. The confusion matrix for per-video classification of boosting on the KTH dataset using 700 codewords is shown in Table 1, which with the increasing of codewords recognition accuracy is rise up at the beginning and if codewords is larger than or equal to 700, the recognition accuracy is stabled. Each cell in the confusion matrix is the average results.

Table 1. Confusion matrix for human action recognition.

	boxing	hand clapping	hand_ waving	walking	jogging	running
boxing	1.00	0.00	0.00	0.00	0.00	0.00
hand clapping	0.18	0.82	0.00	0.00	0.00	0.00
handwaving	0.03	0.02	0.95	0.00	0.00	0.00
walking	0.00	0.04	0.00	0.85	0.10	0.01
jogging	0.00	0.01	0.00	0.15	0.78	0.06
running	0.00	0.00	0.00	0.01	0.07	0.92

As Table 1 show, the average recognition accuracy is stabled to 0.89 when the number of codewords is set as 700. The algorithm correctly classifies most actions. Most of the mistakes the algorithm makes are confusions between "running" and "jogging" actions. This is intuitively reasonable since "running" and "jogging" are similar actions with each other.

To examine the accuracy of our proposed human action recognition approach, we compare our method to two state-of-the-art approaches for human action recognition using the same data and the same experimental settings. The first method is Support Vector Machine (SVM) [23], the second method is LDA [24]. 180 different human action images are used for this experiment. Some images contain the same person but in different actions. The recognition accuracy was observed, which is displayed in Fig. 4.

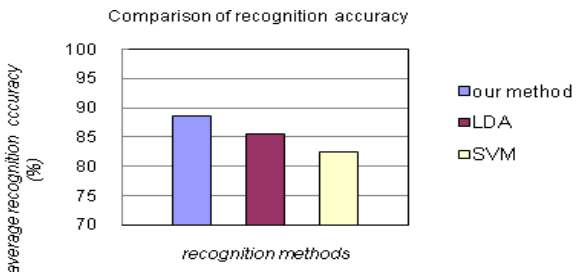

Fig. 4. Comparison of recognition accuracy for three methods.

We can see that our method improves the recognition accuracies. It achieves 88.7 % average recognition rate, whereas "SVM" obtain 82.5 %, "LDA" gets 85.6 %. The reason is that we improve the recognition accuracy in the two stages of human action feature extraction and human action recognition. In the stage of human action feature extraction, we use space-time interest point features and bag-of-words framework that are reliably with noisy image sequences and describe human action effectively. In the stage of human action recognition, we use boosting algorithm to classify human action images. Our method performs significantly better.

6. Conclusions

In this paper, we present a novel method to recognize human action, a "bag-of-words" model combined with a space-time interest point detector and used boosting algorithm for human action categorization. Using a challenging dataset, our experiments validate the proposed model in classification performance. Our algorithm can also recognize multiple actions in complex motion sequences containing multiple actions. Experimental results reveal that our proposed method performs better than previous ones.

Acknowledgements

This work was supported by Research Foundation for Science & Technology Office of Hunan Province under Grant (NO.2012FJ3021) and by the Foundation for Key Constructive Discipline of Hunan Province.

References

[1]. K. Prkachin, S. Mercer, Pain expression in patients with shoulder pathology: validity, coding properties and relation to sickness impact, *Pain*, 39, 1989, pp. 257-265.

[2]. L. Teng, Contextual Bag-of-words for Visual Categorization, *IEEE Transactions on Circuits and Systems for Video Technology*, 21, 4, 2011, pp. 381-392.

[3]. N. Olivera, A. Garg, E. Horvitz, Layered representations for learning and inferring office activity from multiple sensory channels, *Computer Vision and Image Understanding,* 96, 2, 2004, pp. 163-180.

[4]. A. Bobick, A. Wilson, A state-based approach to the representation and recognition of gesture, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19, 12, 1997, pp. 1325-1337.

[5]. T. Xiang, S. Gong, Beyond tracking: modelling activity and understanding behaviour, *Computer Vision*, 67, 1, 2006, pp. 21-51.

[6]. A. Yilmaz, Mubarak Shah, Recognizing human actions in videos acquired by uncalibrated moving cameras, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV' 05)*, 1, 2005, pp. 150-157.

[7]. A. Agarwal, B. Triggs, Learning to track 3d human motion from silhouettes, in *Proceedings of the International Conference on Machine Learning,* Banff, July 2004, pp. 9-16.

[8]. A. Bobick, J. Davis, The recognition of human movement using temporal templates, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23, 3, March 2001, pp. 257-267.

[9]. P. Flaherty, G. Giaever, J. Kumm, *et al.*, A latent variable model for chemogenomic profiling, *Bioinformatics*, 21, 15, 2005, pp. 3286-3293.

[10]. C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local svm approach, *ICPR*, 2004, pp. 32-36.

[11]. V. Cheung, B. Frey, N. Jojic, Video epitomes, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1, 2005, pp. 42-49.

[12]. C. Schmid, R. Mohr, C. Bauckhage, Evaluation of interest point detectors, *International Journal of Computer Vision*, 37, 2, 2000, pp. 151-172.

[13]. I. Laptev, T. Lindeberg, Space-time interest points, in *Proceedings of the 9th IEEE International Conference on Computer Vision*, 1, 2003, pp. 432-439.

[14]. C. Harris, M. Stephens, A combined corner and edge detector, in *Proceedings of the Alvey Vision Conferences*, 1988, pp. 147-52.

[15]. M. Blank, L. Gorelick, E. Shechtman, *et al.*, Actions as space-time shapes, in *Proceedings of the ICCV*, 2005, pp. 1395-1402.

[16]. P. Dollár, V. Rabaud, G. Cottrell, *et al.*, Behavior recognition via sparse spatio-temporal features, in *Proceedings of the VS-PETS*, 2005, pp. 65-72.
[17]. I. Laptev, T. Lindeberg, Space-time interest points, in *Proceedings of the 9th IEEE International Conference on Computer Vision*, Nice, France, 1, 2003, pp. 432-439.
[18]. T. Thi, L. Cheng, J. Zhang, *et al.*, Structured learning of local features for human action classification and localization, *Image and Vision Computing*, 1, 30, 2012, pp. 1-14.
[19]. A. Bosch, X. Munoz, Which is the best way to organize classify images by content, *Image and Vision Computing*, 6, 25, 2007, pp. 778-791.
[20]. Gonzalez A., Perez R., Completeness and consistency conditions for learning fuzzy rules, *Fuzzy Sets and Systems*, 96, 1998, pp. 37-51.
[21]. A Gonzalez, F Herrera, Multi-stage genetic fuzzy systems based on the iterative rule learning approach, *Mathware & Soft Computing*, 4, 1997, pp. 233-249.
[22]. C. Schuldt, L. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in *Proceedings of the IEEE International Conference on Pattern Recognition*, 3, 2004, pp. 32-36.
[23]. P. Dollar, V. Rabaud, G. Cottrell, *et al.*, Behavior recognition via sparse spatio-temporal features, in *Proceedings of the VS-PETS*, 2005, pp. 65-72.
[24]. D. Blei, Y. Andrew, M. Jordan, Latent dirichlet allocation, *Journal of Machine Learning Research*, 3, 2003, pp. 993-1022.