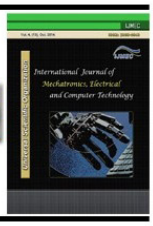
Vol. 6(22), Oct. 2016, PP. 3152-3159

Contents list available at IJMEC

International Journal of Mechatronics, Electrical and Computer Technology (IJMEC)

Journal Homepage: www.aeuso.org

PISSN: 2411-6173 - EISSN: 2305-0543



Human Activity Recognition in Videos Based on a Two Levels K-means and Hierarchical Codebooks

Vahid Hajihashemi[1] and Esmat Pakizeh[2]

[1] *Department of Computer Science, Faculty of Engineering, Kharazmi University, Tehran, Iran*

[2] *Department of Computer Engineering, College of Engineering, Tehran Science and Research Branch, Islamic Azad University, Tehran, Iran*

***Corresponding Author's E-mail:** pakizeh@khu.ac.ir*

Abstract

Human activity recognition is a challenging field in image processing during recent years. Moreover, based on the fast growing the number of videos and multimedia data, there is an enormous demand for an accurate, fast algorithm that analyze, understand and clustered the contents of these data. The speed and accuracy of these methods are so important. Another problem is the large data stored in each video that increase processing, waste time and memory. In This paper a simple novel method for action recognition and video matching based on a clustered codebook model of video frames is presented. Firstly some videos from a single action are analyzed and features related to every frame are extracted. At second step, extracted features of videos are clustered by K-means algorithm and five clusters for each video are made. This step used for decreasing process time and memory needed for saving results. Finally for a single action these clusters merge together and formed a bag of video words (BOV) representation for this action. The proposed method was applied to KTH video dataset for action recognition and the results showed flexibility and correctness of the proposed method in comparison to other methods.

Keywords: *The keywords will separate with commas.*

1. Introduction

Human activity recognition is a challenging field in image processing that used for video surveillance systems, robotics, sports scene detection, gaming and database clustering [1, 2]. According to rapidly growing quantities of videos and multimedia data, there is an enormous demand for an accurate fast algorithm that is capable of analyzing, understanding and clustering these videos' contents. Recognizing human actions in a video is the primary and the most important task of such algorithm. Some unknown parameters such as camera motion, background variations, scale, viewpoint, perspective variations [3–6] affect the algorithm's performance. Also a similar action performed by two persons can appear very different for example running, walking or boxing. In addition clothing, illumination and background changes can increase dissimilarities that lead to decrease accuracy [7–9]. In image processing an action is defined as a human motion performed by a single person in a time about a few seconds, and containing one or more events. Walking, jogging, jumping, running, hand waving, picking up something from the ground, and swimming are some examples of such human simple actions [1,2,6]. The main goal of this paper is recognizing human actions in real environments using a probabilistic video-to-video matching structure. This problem is also referred to as action spotting [10]. Some of current solutions [11, 4] are unable to handle scale variations because the used

Article History:
Received Date: Jun. 14, 2016
Accepted Date: Sep. 11, 2016
Available Online: Oct. 05, 2016

features are too local. To overcome these problems, [13] developed a multi scale hierarchical codebook of BOVs in densely sampled videos which use spatio-temporal compositions. Also [13] is measured similarity between a query and a target dataset using information regarding the most informative spatio-temporal video volumes (STVs). The used features in [13] is static and do not able to model time varying dependencies in an action. For example it cannot distinguish between running backward or forward. In addition the method uses both local and global compositional information of the video frames that need to dense sampling at various scales that is time consuming.

Some algorithms have focused on the action recognition problem by invoking human body models, tracking features or local descriptors [1]. In [16–19] humans' body parts were tracked separately then Codebook Model is made. Clearly, the performance of these algorithms is highly dependent on tracking methods which sometimes do not work correctly for real world video data [20]. Some models try to model the interested points' features based on their movements [21, 22, 19].

Finally in these methods contextual information is computed as a relationship between trajectories [21] or associations between interest points on a trajectory [22]. All of these methods is fully dependent to the tracking method or trajectory models. [23] uses shape template matching for activity recognition, 2D shape matching [23] or its 3D extensions, optical flow matching [13,24,25].

In these methods, single actions are constructed and used to find similar motion patterns. Some methods combined both shape and motion features to achieve better results [26, 27]. In [27], shape and motion descriptors are used simultaneously to find a shape motion pattern for an action. All of mentioned methods require a priori high-level representations of the human motion.

In the proposed method the salient pixels in the video frames are used for finding features in a novel manner that completely differ from conventional background subtraction and salient point detection methods. Proposed method not only optimizes the time of process, but also uses a fully hierarchical training process.

First all training videos analyzed separately, then the results of the first step combined together in order to form the final codebook. The paper is organized as following. In Section 2, Multi-scale hierarchical codebooks and histogram of oriented gradient features are reviewed. The details of proposed method will be explained in Section 3. Experimental results are stated in Section 4. Section 5 states the conclusion and future works.

2. Multi-scale hierarchical codebooks

Each video includes a large number of frames. Each frame is an image and should be processed that lead to enormous number of processing. In order to decrease the process's burden, [31] suggested a semi divide and conquer method that decreases process by mapping each video to some different scales or discarding some of frames (for example in one second of a 30 frame per second video only 10 frames is selected). In these new scales or selection, size or number of images are smaller than the original video that cause to decrease the number of processes. The features that extracted and classified in new frames are defined as a multi scale hierarchical codebook [31].

In order to constructing the multi scale hierarchical codebook, first all frames in a set of videos (query videos) are analyzed and some features are extracted. Then the 3D spatio temporal video volumes are constructed by assuming a matrix of size nx × ny × nt around each pixel (in which nx × ny is the size of the frame images and nt is the length of the video volume in time). Spatio-temporal volume construction is performed at the total playing time of the video that leads to very large number of volumes in the video. Constructed volumes are analyzed and features are assigned to these volumes. In this step the assigned features are histogram of oriented gradients (HOG) based on [31]. Assume that Gx (x,y,t) and Gy (x,y,t) are image gradients and Gt (x,y,t) is the temporal gradient for each pixel at (x,y,t).

$$Gs(x,y,t)=\sqrt{G_x(x,y,t)^2+G_y(x,y,t)^2} \qquad (x,y,t)\in v_i \tag{1}$$

$$\tilde{Gs}(x,y,t)=\frac{Gs(x,y,t)}{\sum_{(x,y,t)\in v_i} Gs(x,y,t)+\varepsilon_{max}} \qquad (x,y,t)\in v_i \tag{2}$$

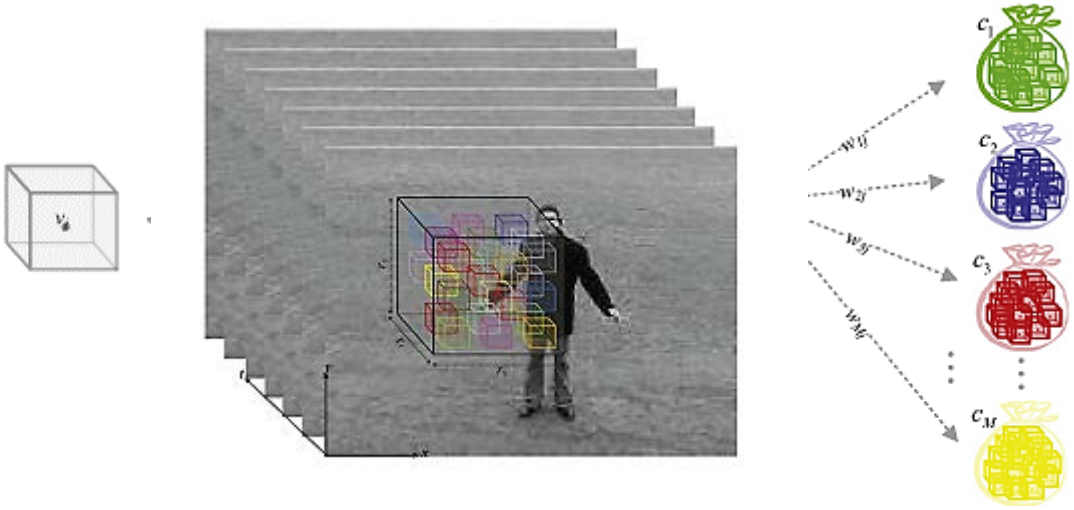


Fig1: each video volume is split to its frames. Codewords define as centers of each volume features. Each codeword is assigned to a volume with a degree of similarity [31]

The spatial gradient used to calculate the normalized 3D gradient in equation 2 to reduce the effect of local texture and contrast. Where $\tilde{Gs}$ the normalized spatial gradient and εmax is a constant, set to 1% of the maximum spatial gradient magnitude in order to avoid numerical instabilities. After this process, the values of Gx (x,y,t), Gy (x,y,t) and Gt (x,y,t) are converted to polar coordinates and the values of M (x,y,t), θ (x,y,t), φ (x,y,t)) are formed where M(x,y,t) is the 3D gradient magnitude, and φ(x,y,t) and θ(x,y, t) are the orientations within [0, π], and [−π, π], respectively.

$$\begin{cases} M(x,y,t)=\sqrt{\tilde{G}_s(x,y,t)^2+G_t(x,y,t)^2} \\ \theta(x,y,t)=arctg\left(\dfrac{G_y(x,y,t)}{G_x(x,y,t)}\right) \\ \theta(x,y,t)=arctg\left(\dfrac{G_t(x,y,t)}{\tilde{G}_s(x,y,t)}\right) \end{cases} \qquad (x,y,t)\in v_i \tag{3}$$

The descriptor vector (HOG) for each video volume has five values (eq. 1 to 3), is calculated using the θ and φ respect to the supposed values for nθ and nφ respectively. Then the HOG is weighted by the gradient magnitude M. The calculated descriptors of each video volume represent motions and their characteristics. Constructed vectors have some degree of robustness to unimportant variations in the data, such as illumination changes [28]. It is obvious that adding some features such as spatio-temporal gradient filters [15], the spatio temporal oriented energy measurements [10] and SIFT1 [29] may be enhance the performance, but in this step for simplicity and dimension reduction we only used HOG features based on [31].

[1] Scale Invariant Feature Transform

3. Proposed method

The main lack of any action recognition method is extremely large number of descriptors that should be analyzed and clustered. Finding any pattern in these large data need a powerful clustering algorithm, large time and enough memory. As the number of these features and vectors is exceptionally large (for example, about 80000×24 based on HOG bins in a one minute video), in the proposed method, these features are hierarchically clustered. In the hierarchically clustering, firstly data is split to some subsets and then the obtained results re-clustered to find final clustering.

Steps of the proposed method are described in the following.

Step1: Constructing low level codebook

After splitting each video into frames and extracting features from each frame, each 10000 frames clustered using k-means with k=5. This is commonly performed in all BOV approaches [30,9,14]. In the proposed method at the first step, similar video volumes were grouped in order to construct the low level codebook. The outcome of this procedure is a set of similar volumes that each of them is known by a center (codeword).

Step2: Constructing public codebook

It is expected that in similar video volumes, these centers are similar to each other so at the next step for a similar action these centers are re-clustered and grouped for constructing a public codebook. This procedure is similar to [15,14] but the re-culstering process is simplified in comparison to [15,14]. Also [31] used a similar process, but its final codebook is so large and complicate. In fact the first step is constructing temporary codewords that is equivalent to the first observed spatio-temporal volume. After that, by re-clustering these codewords, we mapped the local codewords to global codeword area. The number of assigned codewords to each volume in first step is supposed five based on the simulation results. In the other words we compact all 10000 features of each volume to five centers (the number of centers in first step change from 4 to 10 and five is selected based on the best result). After that these groups of five centers re-clustered with kmeans method to a 100 centers (this value is selected based on simulation results similar to first step). For clustering and measuring similarity the Euclidean distance is used in all steps. The output of the final kmeans, last 100 centers, is defined as a public codebook.

Step3: Weighting the center codewords of public codebook

A weight wi,j assigned to each codeword (center) in the public codebook that show the number of samples or centers in the neighborhood of this center. In other words weights show the importance or ratio of each center in the public codebook. The weights are normalized base on equation (4).

$$W_{ij} = \frac{N_{ij}}{N} \tag{4}$$

Where Nij is the number of features that belong to a specified center and N is total number of features.

The main drawback of many BOV2 approaches is that they do not consider the spatio-temporal composition of the video volumes. In the proposed method, with choosing Wij , a probabilistic framework is used for quantifying the spatio-temporal volumes.

[2] Bag of word

The training process is completed at this step. The pseudo code of proposed method is presented for clarity.

Training Algorithm:

Inputs: All database videos with suitable labeling

Output: the K-centers for K category

1. Split video to all frames with an acceptable frame rate
2. Extracting features from all frames on a video based on HOG
3. Constructing Low level Codebook: Find five center for each video (if K-means in this step do not converge, split frames to some groups and repeat from step 2), store these centers
4. Repeat steps 2-3 for all videos
5. Constructing a public codebook: Combining each category centers with performing K-means in order to reach 100 centers

Testing Algorithm:

Inputs: All category centers (one hundred for each type) found in training process, the new video

Output: the category of new video

1. Split new video to all frames with an acceptable frame rate
2. Extracting features from all frames based on HOG
3. Find five centers for new video (if K-means in this step do not converge, split frames to some groups and repeat from step 2) and store these centers
4. Compare these centers with all of category centers found in step 5 of training process
5. Assign a membership probability to each category (weighting summation)
6. Choosing the maximum score as new video category

After the new video is analyzed, the goal is to measure the likelihood of this video to the target videos given the query. To specify this, it is necessary to analyze the similarity of this video codewords to those stored in each action codebook. Firstly each codeword of new video, is compare to all codewords belong to codebook of an action. The maximum similarity is calculated based on minimum Euclidean distance. This value is then multiple with Wij of chosen codeword. At last all of this value is added each other and give a weight for similarity of the new video to target videos. This process is repeated for all codebooks (actions) and finally the maximum similarity is chose based on computed weights.

4. Experimental results

In order to examine the proposed method's performance, it was tested on the KTH dataset [12]. The KTH dataset is one of the standard benchmarks in the literature that contains six different actions including (boxing, running, walking, hand waving, hand clapping, jogging), performed by twenty five

different persons in four different scenarios (indoor, outdoor, outdoor at different scales, outdoor with different clothes). The proposed method is tested with about 50% of dataset videos for training and remain videos for testing. The results are showed in table 1. The results confirm the capability of proposed method in video matching.

Table 1: The result of proposed method in KTH dataset

	boxing	**handclapping**	**jogging**	**Running**	**walking**	**handwaving**
boxing	82.82%	8.08%	1.04%	1.01%	3.03%	4.02%
Handclapping	7.07%	83.47%	3.03%	2.05%	1.07%	3.30%
jogging	1.00%	2.00%	85.00%	6.00%	4.00%	2.00%
running	1.00%	0.00%	7.00%	87.00%	3.00%	1.00%
walking	0.00%	1.00%	9.00%	5.00%	83.00%	2.00%
handwaving	4.00%	5.00%	2.00%	1.00%	4.00%	84.00%

The table 1 shows that defined codebooks is exactly found the main type of action. The KTH dataset has two main category, Actions with hand including (boxing, hand waving and handclapping) and actions with foot including (running, jogging and walking), the proposed method determine two types so good but in a one category some mismatches have happened. For example in the boxing row, the proposed method has about 83% correctness, the main false detection is for hand clapping and hand waving that belong two hand category actions and similar to boxing. Also the proposed method is compared to the method of [31]. The results of examining the method of [31] with a limited query video are presented in table2.

As showed in the two tables, the proposed method is better than method of [31] in recognizing last four actions (jogging, running, walking and handwaving). Also it is approximately comparable with method of [31] in two first action (boxing, clapping). Due to the simplicity and levels of proposed method, the complexity, runtime and needed memory for our method is lower than the [31,14,15] methods. If the time step for discarding frames and other parameters supposed equal in proposed method and [31], the proposed method is about 50% better in runtime and is similar in memory to [31].

Table 2: The result of [31] with a limited query in KTH dataset. (Fig 6. b)

	boxing	**handclapping**	**jogging**	**running**	**walking**	**handwaving**
boxing	0.86	0.04	0.07	0.02	0	0.01
Clapping	0.05	0.84	0.07	0.02	0.02	0
Waving	0.06	0.11	0.81	0.01	0	0.01
Jogging	0.03	0.01	0	0.79	0.11	0.06
Running	0.02	0.03	0	0.12	0.75	0.08
Walking	0.03	0	0.02	0.07	0.06	0.8

Conclusion and Future works

In this paper a simple novel method for action recognition and video matching based on kmeans algorithm and a clustered codebook of features is proposed. In the hierarchical proposed method firstly some videos is analyzed separately to extract temporary features, these features is clustered by kmeans in at least five neighborhood or center. At second step, the centers of videos belonging to an action is reclustered by K nearest neighborhood method and a codebook with 100 centers (codeword) is made. In the meantime a weight is assigned to each codeword due to its importance in codebook. Based on these codebooks, a new video is analyzed by checking similarity of its codewords to stored codewords in all action codebooks. The weight of each codeword is used in this step and in fact a weighted mean is computed based on stored weights to enhance correctness of proposed method. The proposed method does not require any prior information about actions, background, scene, etc, each video is analyzed separately and only five centers of each video need to be stored for next step. This process is help to saving memory and simplify training process. Using markov chains as a time dependent sequential decision making algorithm maybe enhance correctness of the proposed method. Finally the proposed method was applied to KTH video dataset for action recognition and the results showed flexibility, correctness and lower runtime of the proposed method in comparison to other methods.

References

[1] R. Poppe, A survey on vision-based human action recognition, Image Vision Comput. 28 (6) (2010) 976–990.

[2] P. Turaga, R. Chellappa, V.S. Subrahmanian, O. Udrea,Machine recognition of human activities: a survey, IEEE Trans. Circuits Syst. Video Technol. 18 (11) (2008) 1473–1488.

[3] Ronald Poppe, A survey on vision-based human action recognition, Image and Vision Computing, Volume 28, Issue 6, June 2010, Pages 976-990.

[4] S. Savarese, A. DelPozo, J.C. Niebles, F.-F. Li, Spatial–temporal correlations for unsupervised action classification, WMVC, 2008, pp. 1–8.

[5] L. Wang, L. Cheng, Elastic sequence correlation for human action analysis, IEEE Trans. Image Process. 20 (6) (2011) 1725–1738.

[6] D.Weinland, R. Ronfard, E. Boyer, A survey of vision-based methods for action representation, segmentation and recognition, Comput. Vision Image Underst. 115 (2) (2011) 224–241.

[7] Thomas B. Moeslund, Adrian Hilton, Volker Krüger, A survey of advances in vision-based human motion capture and analysis, Computer Vision and Image Understanding, Volume 104, Issues 2–3, November–December 2006, Pages 90-126.

[8] K. Mikolajczyk, H. Uemura, Action recognition with appearance–motion features and fast search trees, Comput. Vision Image Underst. 115 (3) (2011) 426–438.

[9] H. Seo, P. Milanfar, Action recognition from one example, IEEE Trans. Pattern Anal. Mach. Intell. 33 (5) (2011) 867–882.

[10] Sadanand, S.; Corso, J.J., "Action bank: A high-level representation of activity in video," Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on , vol., no., pp.1234,1241, 16-21 June 2012.

[11] A. Oikonomopoulos, I. Patras, M. Pantic, Spatiotemporal localization and categorization of human actions in unsegmented image sequences, IEEE Trans. Image Process. 20 (4) (2011) 1126–1140.

[12] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, ICPR, vol. 3, 2004, pp. 32–36.

[13] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, IEEE Trans. Pattern Anal. Mach. Intell. 29 (12) (2007) 2247–2253.

[14] M. Javan Roshtkhari, M.D. Levine, An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions, Comput. Vision Image Underst. 117 (10) (2013) 1436–1452.

[15] Weiming Hu; Xuejuan Xiao; Zhouyu Fu; Xie, D.; Tieniu Tan; Maybank, S., "A system for learning statistical motion patterns," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.28, no.9, pp.1450,1464, Sept. 2006.

[16] [16] Jhuang, H.; Serre, T.; Wolf, L.; Poggio, T., "A Biologically Inspired System for Action Recognition," Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, vol., no., pp.1, 8, 14-21 Oct. 2007.

[17] C. Rao, A. Yilmaz,M. Shah, View-invariant representation and recognition of actions, Int. J. Comput. Vision 50 (2) (2002) 203–226.

[18] F. Yuan, G.-S. Xia, H. Sahbi, V. Prinet, Mid-level features and spatio-temporal context for activity recognition, Pattern Recogn. 45 (12) (2012) 4182–4191.

[19] H. Wang, A. Kl¨ aser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. IJCV, 103(1):60–79, 2013.

[20] M.J. Roshtkhari, M.D. Levine, Online dominant and anomalous behavior detection in videos, Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, 2013, pp. 2609–2616.

[21] R. Messing, C. Pal, H. Kautz, Activity recognition using the velocity histories of tracked keypoints, Computer Vision, 2009 IEEE 12th International Conference on, 2009, pp. 104–111.

[22] Kovashka, A.; Grauman, K., "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, vol., no., pp.2046, 2053, 13-18 June 2010.

[23] A. Yilmaz, M. Shah, Actions sketch: a novel action representation, Computer Vision and Pattern Recognition (CVPR), IEEE Conference on, 2005, pp. 984–989.

[24] E. Shechtman, M. Irani, Space-time behavior-based correlation-or-how to tell if two underlying motion fields are similar without computing them? IEEE Trans. Pattern Anal. Mach. Intell. 29 (11) (2007) 2045–2056.

[25] Dollar, P.; Rabaud, V.; Cottrell, G.; Belongie, S., "Behavior recognition via sparse spatio-temporal features," Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on, vol., no., pp.65, 72, 15-16 Oct. 2005.

[26] Y. Ke, R. Sukthankar, M. Hebert, Volumetric features for video event detection, Int. J. Comput. Vision 88 (3) (2010) 339–362.

[27] Z. Jiang, L. Zhe, L.S. Davis, Recognizing human actions by learning and matching shape–motion prototype trees, IEEE Trans. Pattern Anal. Mach. Intell. 34 (3) (2012) 533–547.

[28] M. Bertini, A. Del Bimbo, L. Seidenari, Multi-scale and real-time non-parametric approach for anomaly detection and localization, Comput. Vision Image Underst. 116 (3) (2012) 320–329.

[29] Ali, S.; Shah, M., "Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.32, no.2, pp.288,303, Feb. 2010.

[30] Jingen Liu; Jiebo Luo; Shah, M., "Recognizing realistic actions from videos “in the wild”," Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, vol., no., pp.1996, 2003, 20-25 June 2009.

[31] Mehrsan Javan Roshtkhari, Martin D. Levine, Human activity recognition in videos using a single example, Image and Vision Computing, Vol. 31, no 11, November 2013, pp. 864-876

Authors

Vahid Hajihashemi is currently a master student of Computer Engineering at Faculty of Engineering at Kharazmi University of Tehran. His research interests include artificial intelligence and evolutionary computations.

Esmat Pakizeh received her M.S in Artificial Intelligence from Isfahan University of Technology, Isfahan, Iran, in 2010, her B.S degree in software engineering from Kharazmi University, Tehran, Iran, in 2007. Now, she is a PhD candidate of Artificial Intelligence at Islamic Azad University, Science and Research Branch, Tehran, Iran. Also she joined to Cognitive Science Laboratory at Kharazmi University in 2012. Her research interests include evolutionary algorithms, multiagent learning, data mining, reinforcement learning and neural networks.