ORIGINAL ARTICLE

An enhanced method for human action recognition

Mona M. Moussa a,*, Elsayed Hamayed b, Magda B. Fayek b, Heba A. El Nemr a

a Computers and Systems Department, Electronics Research Institute, Egypt
b Computer Engineering Department, Faculty of Engineering, Cairo University, Egypt

ARTICLE INFO

Article history:
Received 28 July 2013
Received in revised form 26 November 2013
Accepted 27 November 2013
Available online 5 December 2013

Keywords:
SIFT
Action recognition
Bag of words
SVM

ABSTRACT

This paper presents a fast and simple method for human action recognition. The proposed technique relies on detecting interest points using SIFT (scale invariant feature transform) from each frame of the video. A fine-tuning step is used here to limit the number of interesting points according to the amount of details. Then the popular approach Bag of Video Words is applied with a new normalization technique. This normalization technique remarkably improves the results. Finally a multi class linear Support Vector Machine (SVM) is utilized for classification. Experiments were conducted on the KTH and Weizmann datasets. The results demonstrate that our approach outperforms most existing methods, achieving accuracy of 97.89% for KTH and 96.66% for Weizmann.

© 2013 Production and hosting by Elsevier B.V. on behalf of Cairo University.

Introduction

Human action recognition is an active area of research due to the wide applications depending on it as detecting certain activities in surveillance video, automatic video indexing and retrieval, and content based video retrieval.

Action representation can be categorized as: flow based approaches [1], spatio-temporal shape template based approaches [2,3], tracking based approaches [4] and interest points based approaches [5]. In flow based approaches optical flow computation is used to describe motion, it is sensitive to noise and cannot reveal the true motions. Spatio-temporal shape template based approaches treat the action recognition problem as a 3D object recognition problem and extracts features from the 3D volume. The extracted features are very huge so the computational cost is unacceptable for real-time applications. Tracking based approaches suffer from the same problems. Interest points based approaches have the advantage of short feature vectors; hence low computational cost. They are widely used and are adopted in this work.

One of the widely used techniques in the action recognition task is Bag of Video Words (BoVW) [6]; which is inspired from bag of words model in natural language processing, where videos are treated as documents and visual features as words [7,8]. This approach proved its robustness to location changes and to noise. Usually the system consists of four main steps: interest-points detection, features description, vector quantization and normalization of the features to construct histogram representation. Finally the histograms are used for classification.

In this work SIFT [9] is used for detecting interest points where the extracted features are invariant to scale, location and orientation changes. 2D SIFT has another advantage which is the limited size of the features vectors; which consumes less computation time than other techniques such as...
Global descriptors that jointly encode shape and motion were suggested by Lin et al. [10], while Liu and Shah [11] suggested a method to automatically find the optimal number of visual word clusters through maximization of mutual information (MMI) between words and actions. MMI clustering is used after k-means to discover a compact representation from the initial codebook of words. They showed some performance improvement.

Bregonzio et al. [12] exploited only the global distribution information of interest points. In particular, holistic features from clouds of interest points accumulated over multiple temporal scales are extracted. A feature fusion method is formulated based on Multiple Kernel Learning.

Chen and Hauptmann [5] proposed MoSIFT which detects interest points then encodes their local appearance and models the local motion. First the well-known SIFT algorithm is applied to find visually distinctive components in the spatial domain and detect spatio-temporal interest points with (temporal) motion constraints. The motion constraint consists of a 'sufficient' amount of optical flow around the distinctive points.

Niebles et al. [13] used probabilistic Latent Semantic Analysis (pLSA) model and Latent Dirichlet Allocation (LDA) to automatically learn the probability distributions of the spatial–temporal words and the intermediate topics corresponding to human action categories. The system can recognize and localize multiple actions in long and complex video sequences containing multiple motions.

Sadanand and Corso [14] presents a high-level representation of video where individual detectors in this action bank capture example actions, such as “running-left” and “biking-away,” and are run at multiple scales over the input video; it represents a video as the collected output of many action detectors that each produces a correlation volume. Being a template-based method, there is actually no training of the individual bank detectors, the detector templates in the bank are selected manually. This method requires using a number of action templates as detectors, which is compositionally expensive in practice.

Tran et al. [15] combined both local and global representations of the human body parts, encoding the relevant motion information as well as being robust to local appearance changes. It represented motion of body parts in a sparse quantized polar space as the activity descriptor.

Fathi and Mori [1] constructed a mid-level motion features built from low-level optical flow information (which is sensitive to noise). These features are focused on local regions of the image sequence, computed on a figure-centric representation, and are created using a variant of AdaBoost. Mid-level shape features were constructed from low-level gradient features using also the AdaBoost algorithm.

Kovashka and Grauman [16] first extract local motion and appearance features from training videos, quantizes them to a visual vocabulary, and then forms candidate neighborhoods consisting of the words associated with nearby points and their orientation with respect to the central interest point. Detectors for these variable-sized neighborhoods are then recursively mapped to higher-level vocabularies, producing a hierarchy of space–time configurations at successively broader scales.

**Methodology**

The proposed system is composed of four stages (as shown in Fig. 1): detection of interesting points, feature description for the detected points, building the codebook and finally the classification.

**Enhanced interesting points detection**

First step in the system is interest points detection where SIFT is utilized to do this process, using algorithm [17]. Fine tuning the threshold parameter is performed to adjust the number of interest points automatically according to the amount of details in each frame. The fine tuning is done by initially apply threshold value = 6 then according to the number of extracted interesting points (np) the threshold (th) is set to a new value as follows:

- if np > 25 then th = 14
- else if np > 20 then th = 10
- else if np > 10 then th = 8
- else th = 6

The threshold value determines the amount of details the detector returns, so when the threshold value is high only the important interest points are detected, while the weak interest points are neglected. Thus the useful information is not lost.

Fig. 2 shows the enhancement achieved by adjusting the threshold. It is obvious that without using a threshold the number of extracted points is very high and they are insignificant where most of them lied in the background. Utilizing a threshold, only the significant points are detected without the need for an additional segmentation step which represents significant processing overhead.

**Features description**

The SIFT feature vector consists of 128 elements, the coordinates of each point (the x and y location in the frame) are
made use of to enhance the results as inspired by Lai et al. [18], so the new feature vector becomes 130 elements (the old 128 elements vector + x coordinate of the interest point + y coordinate of the interest point). One of the reasons to use SIFT (beside that it is invariant to scale, location and orientation changes) is its short feature vector which does not need to use topic modeling methods as pLSA and LDA, where a separate topic model is learned for each action class and new samples are classified by using the constructed action topic models.

Building and normalizing the codebook

After feature extraction the next step is building the codebook where \( K \)-means [19] clustering algorithm is utilized. The \( K \)-means clustering is the most popular method to construct visual dictionary due to its simplicity and speed of convergence. \( K \)-means use the generated descriptors of the interest points to cluster them; the resulted clusters centers are called visual words, and the word vocabulary is the set of these words. Then the descriptors are mapped to the vocabulary to build a word frequency histogram, so each video has a signature which is a histogram that reflects the words frequency in it.

A similar method as Niebles et al. [13] is followed for the KTH dataset, since the total number of features from all training examples is very large to use for clustering, only videos of two actors are used to learn the codebook. The codebook size was examined to have values ranging from 900 to 1300 for KTH dataset. Fig. 3 demonstrates the effect of changing the codebook size on the results accuracy. The results indicate that the best accuracy is achieved with a code book size of 1100. For the Weizmann dataset all the training set is used to build the codebook with size 200.

To deal with actions with variable durations, the histograms representing the videos need to be normalized to ensure that the resulting histograms have the same dimension. Wang et al. [20] reviewed three methods for normalization:

\[ p = \frac{p}{\sum_{k=1}^{K} p_k} \]  
\[ p = \sqrt{\frac{1}{\sum_{k=1}^{K} p_k^2}} \]  
\[ f(p_k) = \text{sign}(p_k)|p_k|^\alpha \]

where \( p \) is the histogram to be normalized, \( p_k \) is one of its components and \( 0 < \alpha < 1 \) is a parameter for normalization.

In this work min–max normalization [21] technique is used; which is one of the famous techniques used for data normalization; to normalize the data from zero to one. In this method all the histograms to be normalized are treated as one two-dimensional matrix, the rows represent the videos and the columns represent the histograms bins. Normalization is then applied on each column using the following equation:

\[ p_{ij} = \frac{p_{ij} - \min(p_i)}{\max(p_i) - \min(p_i)} \]

where \( p_{ij} \) is the value of bin number \( j \) to be normalized in video number \( i \), \( \max(p_i) \) and \( \min(p_i) \) are the maximum and minimum values respectively in bin \( j \) over all the videos, now all values are between 0 and 1.

Classification

Here comes the SVM role for classification. In machine learning SVM is a supervised learning model with associated learning algorithms that analyze data and recognize patterns. An
SVM model is a representation of the examples as points in space. Given a set of training examples, each marked as belonging to one of the categories, SVM maps them so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

A linear multi class SVM [22] is trained using the normalized histograms. In the testing step, the training histograms are re-normalized along with the one for testing. The re-normalization step is done so that the resultant normalized test histogram is affected by all the histograms (training ones and testing one). Afterward, the resultant normalized test histogram is fed to the SVM to be classified.

**Results and discussion**

Due to the limited number of samples (persons) in the dataset, the leave-one-out method has been adopted [23] where each run uses 24 persons (videos) for clustering and training and one person for testing. Then the average is calculated to give the final recognition rate. Thus, in this work leave-one-person-out is used for KTH and Weizmann datasets and this work is compared mainly with the others using the same setup.

**Using KTH dataset**

KTH dataset was provided by Schuldt et al. [6] in 2004 and is one of the largest public human activity video dataset, it consists of six action class (boxing, hand clapping, hand waving, jogging, running and walking) each action is performed by 25 actors each of them in four different scenarios including indoor, outdoor, changes in clothing and variations in scale.

As mentioned above leave-one-person-out experimental setup is used in this work, where each run uses 24 persons for clustering and training, and one person for testing (24 videos). Then, the average of the results is computed to be the final result.

Table 1a–d present the confusion matrices of KTH dataset using $\ell$1-Normalization, $\ell$2-Normalization, power-Normalization and the proposed normalization technique respectively. The recognition results are presented in the form of average recognition rates. Each entry in the table gives the rate of recognizing the row action (ground truth) by the column action. Table 1e presents the accuracy using the proposed method for each of the four scenarios (outdoor, variations in scale, changes in clothing and indoor). Table 2 presents a comparison between the overall results (recognition rate) achieved using these normalization methods and also a combination of

| Table 1(a) | Confusion matrix of KTH dataset using $\ell$1-Normalization. |
|---|---|
| Boxing | Clapping | Waving | Jogging | Running | Walking |
| Boxing | 0.2 | 0.59 | 0.2 | 0 | 0 | 0.01 |
| Clapping | 0.03 | 0.91 | 0.03 | 0.01 | 0.02 | 0 |
| Waving | 0.02 | 0.2 | 0.76 | 0.02 | 0 | 0 |
| Jogging | 0 | 0 | 0.04 | 0.56 | 0.28 | 0.12 |
| Running | 0 | 0 | 0 | 0.16 | 0.63 | 0.21 |
| Walking | 0 | 0 | 0 | 0.13 | 0.29 | 0.58 |

| Table 1(b) | Confusion matrix of KTH dataset using $\ell$2-Normalization. |
|---|---|
| Boxing | Clapping | Waving | Jogging | Running | Walking |
| Boxing | 0.55 | 0.39 | 0.04 | 0.02 | 0 | 0 |
| Clapping | 0.14 | 0.81 | 0.05 | 0 | 0 | 0 |
| Waving | 0.04 | 0.21 | 0.74 | 0.01 | 0 | 0 |
| Jogging | 0 | 0 | 0.01 | 0.6 | 0.27 | 0.12 |
| Running | 0 | 0 | 0 | 0.26 | 0.6 | 0.14 |
| Walking | 0 | 0 | 0 | 0.16 | 0.15 | 0.69 |

| Table 1(c) | Confusion matrix of KTH dataset using power-normalization. |
|---|---|
| Boxing | Clapping | Waving | Jogging | Running | Walking |
| Boxing | 0.99 | 0.01 | 0 | 0 | 0 | 0 |
| Clapping | 0.04 | 0.92 | 0.04 | 0 | 0 | 0 |
| Waving | 0 | 0.04 | 0.96 | 0 | 0 | 0 |
| Jogging | 0 | 0 | 0.02 | 0.98 | 0 | 0 |
| Running | 0 | 0 | 0 | 0.02 | 0.98 | 0 |
| Walking | 0 | 0 | 0 | 0 | 0.03 | 0.97 |
them. As shown the proposed normalization technique proved positive effort on the performance, and it is worth mentioning that most of the wrong classified actions were done by the same actor.

Table 2 also shows the effect of each normalization technique on the processing time (time taken to calculate it + time needed for SVM to train and test). As can be noticed, the proposed normalization takes (about 2.5 s) more than the time needed for power normalization (the fastest one) for the 25 runs. So time is increased slightly in some cases versus a good improvement in accuracy in all cases.

Table 3 shows a comparison between our method and a group of other previously proposed systems that use leave-one-out setup. The results show that for the KTH dataset our result is the best of them.

Table 4 Confusion matrix of Weizmann dataset.
Using Weizmann dataset

Weizmann dataset is introduced by Blank [2] in 2005, it consists of 10 actions: bending, jumping jack, jumping, jumping in place, running, galloping sideways, skipping, walking, one-hand-waving and two-hands-waving. Each of these actions is performed by 9 actors resulting in 90 videos.

Leave-one-person out experimental setup is also used with the Weizmann dataset; where at each run 8 persons are used for clustering and training, and one person for testing (10 videos). Then the average of the results is taken as a measure of accuracy. Table 4 shows the confusion matrix of the Weizmann dataset, where most of the actions are classified correctly and the ones that are classified wrong are only three videos out of the 90 videos.

For the Weizmann dataset our result (Table 3) is the second best one. Lin et al. [10] combines shape and motion descriptors, with accuracy 81.11% for using shape only descriptor and with accuracy 88.89% for motion only descriptor. While the accuracy of 100% is achieved by combining both, this increases the processing time. The method proposed by Fathi and Mori [1] is based on action templates which cannot represent variations in time, speed, and action style through special variables. Variations are instead implicitly represented through large sets of example sequences. So they proposed an advanced statistical learning method “Adaboost”, making the classification problem more difficult.

Conclusions

This work presents a human action recognition system that is fast and simple. The system is composed of four stages: detection of interesting points, features description, the bag of visual words, and classification. For the first and second steps SIFT is used, the traditional k-means clustering is utilized to build the BoVW, and finally multi class linear SVM is employed for classification. The proposed normalization method as well as the adjustment of the threshold value for SIFT has enhanced the result of detection of the interesting points (by 2%) comparing to other systems.

Future work includes applying the proposed system on different complex datasets, such as: sports and real actions ones. These datasets are more complex than the ones used here and the system may need some improvements to achieve acceptable recognition rate. Also the use of a sequence of different actions to segment it then recognize each action is another point of research in the future work.

Conflict of interest

The authors have declared no conflict of interest.

Compliance with Ethics Requirements

This article does not contain any studies with human or animal subjects.

References

[1] Fathi A, Mori G. Action recognition by learning mid-level motion features. Comput Vision Pattern Recogn, CVPR IEEE 2008:1–8.
[2] Blank M, Gorelick L, Shechtman E, Irani M, Basri R. Actions as space-time shapes. Int Conf Comput Vision, ICCV IEEE 2005;2:1395–402.
[3] Ke Y, Sukthankar R, Hebert M. Efficient visual event detection using volumetric features. Int Conf Comput Vision, ICCV IEEE 2005;1:166–73.
[4] Sheikh Y, Sheikh M, Shah M. Exploring the space of a human action. Int Conf Comput Vision, ICCV IEEE 2005:144–9.
[5] Chen MY, Hauptmann AG. MoSIFT: recognizing human actions in surveillance videos. Technological report, CMU-CS-09-161, Carnegie Mellon University; 2009. p. 9–161.
[6] Schuld C, Laptev I, Caputo B. Recognizing human actions: a local SVM approach. Int Conf Pattern Recogn, ICPR IEEE 2004;3:32–6.
[7] Csurka G, Dance C, Fan L, Willamowski J, Bray C. Visual categorization with bags of key points. ECCV International Workshop on Statistical Learning in Computer Vision 2004:1–22.
[8] Gemert J, Geusebroeck J, Veenman C, Smeulders A. Kernel codebooks for scene categorization. Proc Euro Conf Comput Vision, ECCV 2008:696–709.
[9] Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vision 2004;60(2):91–110.
[10] Lin Z, Jiang Z, Davis LS. Recognizing actions by shape-motion prototype trees. Int Comput Vision, ICCV IEEE. p. 1–8.
[11] Liu J, Shah M. Learning human actions via information maximization. Comput Vision Pattern Recogn, CVPR IEEE 2008:1–8.
[12] Bregonzio M, Xiang T, Gong S. Fusing appearance and distribution information of interest points for action recognition. Pattern Recogn 2012;45(3):1220–34.
[13] Niebles J, Wang H, Fei-Fei L. Unsupervised learning of human action categories using spatial-temporal words. Int J Comput Vision 2008;79(3):299–318.
[14] Sadanand S, Corso J. Action bank: a high-level representation of activity in video. Comput Vision Pattern Recogn, CVPR IEEE 2012:1234–41.
[15] Tran KN, Kakadiaris IA, Shah SK. Modeling motion of body parts for action recognition. British Mach Vision Conf, BMVC 2011.
[16] Kovashka A, Grauman K. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. Comput Vision Pattern Recogn, CVPR IEEE 2010:2046–53.
[17] Vedaldi A, Fulkerson B, VLFeat. An open and portable library of computer vision algorithms; 2008. <http://www.vlfeat.org/>.
[18] Lai KT, Hsieh CH, Lai MF, Chen MS. Human action recognition using key points displacement. Int Conf Image Signal Process, ICISP 2010:6134:439–47.
[19] MacQueen JB. Some methods for classification and analysis of multivariate observations. Proc 5th Berkeley symposium on mathematical statistics and probability 1967:1:281–97.
[20] Wang X, Wang L, Qiao Y. Comparative study of encoding, pooling and normalization methods for action recognition. Asian Conf Comput Vision, ACCV 2012;7726:572–85.
[21] Jayalakshmi T, Santhakumaran A. Statistical normalization and back propagation for classification. Int J Comput Theor Eng (IJCTE) 2011;3(1):89–93.
[22] Chang C, Lin C. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol, TIST 2011;2(3):1–27.
[23] Gao Z, Chen MY, Hauptmann AG, Cai A. Comparing evaluation protocols on the KTH dataset. In: International conference on human behavior understanding, vol. 6219, Springer; 2010. p. 88–100.
[24] Cao L, Liu Z, Huang TS. Cross-dataset action detection. Comput Vision Pattern Recog, CVPR IEEE 2010:1998–2005.
[25] Kaaniche MB, Bremond F. Gesture recognition by learning local motion signatures. Comput Vision Pattern Recog, CVPR IEEE 2010:2745–52.
[26] Dollar P, Rabaud V, Cottrell G, Belongie S. Behavior recognition via sparse spatio-temporal features. IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance 2005:65–72.
[27] Klaser A, Marszaek M, Schmid C. A spatio-temporal descriptor based on 3D-gradients. British Mach Vision Conf, BMVC 2008:995–1004.
[28] Zhang Z, Hu Y, Chan S, Chia LT. Motion context: a new representation for human action recognition. Proceedings of the European conference on computer vision, ECCV Springer 2008;5305:817–29.