Fuzzy-Weighted Efficient and Robust Human Action Video Retrieval Postprint

Authors: Zhang Han, Han Yi, Li Yuexin

Date: 2018-05-18T00:00:00+00:00

Abstract

To improve the robustness and efficiency of human action video retrieval, a fuzzy-weighted human action video retrieval method is proposed. This method employs the 3D Harris operator to detect spatio-temporal interest points in videos, extracts gradient information from these interest points, and constructs feature vectors; then employs a fuzzy clustering method to construct clustered feature vectors, enhancing the anti-interference capability of the feature vectors; subsequently matches gradient vector pairs in the clustered feature vectors, constructs a fuzzy weight matrix, and calculates the similarity between the query video and each video in the database; finally, conducts video retrieval experiments on the KTH dataset, and evaluates the performance using three metrics: precision, recall, and retrieval time, thereby demonstrating the superior performance of the proposed method.

Full Text

Preamble

Title: Efficient and Robust Video Retrieval for Human Activity with Fuzzy Weighting

Authors: Zhang Han[1], Han Yi[1,2], Li Yuexin[3]
[1]College of Computer Science & Information Engineering, Anyang Institute of Technology, Anyang, Henan 455000, China
[2]National NC System Engineering Research Center, Huazhong University of Science & Technology, Wuhan 430000, China
[3]School of Computer Science & Engineering, Hubei University, Wuhan 430000, China

Abstract: To improve the robustness and efficiency of human activity video retrieval, this paper proposes a fuzzy-weighted approach for retrieving human

activity videos. The method employs the 3D Harris operator to detect spatio-temporal interest points in videos and extracts gradient information from these points to construct feature vectors. Fuzzy clustering is then applied to build clustered feature vectors, enhancing their anti-interference capability. Subsequently, gradient vector pairs in the clustered feature vectors are matched to construct a fuzzy weight matrix, which calculates the similarity between a query video and each video in the database. Finally, video retrieval experiments conducted on the KTH database demonstrate the method' s superior performance when evaluated using three metrics: precision, recall, and retrieval time.

Keywords: video retrieval; behavior recognition; fuzzy clustering; spatio-temporal interest points; 3D Harris

0 Introduction

Human activity video retrieval, though based on human activity recognition technology, differs significantly from it. The primary distinction lies in the fact that video retrieval uses only a single training sample—the query video itself. Consequently, unlike activity recognition, which can learn from numerous samples to build classifiers, retrieval must infer correlations between videos solely through similarity computation, making the task inherently more challenging. Moreover, video retrieval systems generally demand high efficiency, necessitating that algorithm design also consider timeliness.

With the exponential growth of text, image, and video data driven by internet and multimedia technologies, video retrieval has emerged as an increasingly important research direction. While text and image retrieval have matured over many years, video retrieval remains relatively nascent. Existing methods in human activity video retrieval still lack satisfactory robustness and efficiency. For instance, reference [5] proposes a fast feature correspondence method to compute matching costs as a similarity metric, embedding this within a multi-ranking framework for action retrieval. Reference [6] introduces a relevance feedback video retrieval system using SVM with active learning to improve performance. Reference [7] models video context information using semi-supervised learning paradigms to effectively model action patterns from few examples. However, these approaches exhibit limited robustness and computational efficiency.

To address these limitations, this paper proposes a fuzzy-weighted method for human activity video retrieval. The primary contributions include: (1) introducing fuzzy clustering to construct clustered feature vectors, thereby enhancing anti-interference capability; and (2) building a fuzzy weight matrix to modify Euclidean distance measures, improving the robustness of similarity computation. The proposed method also features low computational complexity and high efficiency.

1 Overview of Human Activity Recognition

Human activity recognition is a prominent research area in computer vision that aims to automatically analyze and categorize ongoing human actions from unknown videos or image sequences. A typical recognition framework [8,9] is illustrated in [Figure 1: see original paper].

The process begins with motion target detection to quickly extract regions of interest from action videos or images, reducing the difficulty of subsequent feature extraction and classification. Established techniques such as frame differencing, background subtraction, and optical flow can be employed. Additionally, human detection methods (e.g., HOG features with SVM classifiers) can identify whether motion regions contain human figures, eliminating non-human regions from further processing. However, motion detection is not mandatory, as some recognition algorithms extract features from entire images rather than focusing on moving human targets.

Feature extraction constitutes the core of human activity recognition. Effective features must not only capture distinctions between different actions but also accommodate variations within the same action category. Common features include silhouette, optical flow, gradient, spatio-temporal, and depth features. Silhouette features, constructed from boundary points or shape contexts (e.g., silhouette energy images, shape-from-silhouette), are robust to color and texture but sensitive to occlusion and incomplete contour extraction. Optical flow captures instantaneous pixel motion, reflecting speed and direction, yet is affected by imaging conditions (e.g., camera distance, field of view) and suffers from low computational efficiency. Spatio-temporal features, such as spatio-temporal interest points and context features, are most widely used. Interest points describe actions through isolated point features detected by operators like 3D-Harris, SIFT, and Dollar, followed by descriptor extraction (e.g., HOG, space-time local regression kernels). Context features model relationships between actions and environments through scene, spatial, and scale contexts, often using Markov logic networks. Depth features incorporate spatial depth information but require specialized acquisition equipment, limiting their applicability.

Action understanding and classification compares extracted features against learned prior knowledge through pattern recognition. Models include human body models (2D stick/ribbon models, 3D conical models) that analyze model variations, and statistical models (spatio-temporal templates, dynamic programming, state transition models) that statistically analyze feature variations. Classification employs machine learning methods such as SVM, random forests, and neural networks to construct action classifiers.

2 Human Activity Video Retrieval

Given a query video, a human activity video retrieval algorithm must search a video database for relevant videos based on extracted activity features. The algorithm comprises two core components: (1) describing a video as a feature

vector that characterizes an action class while distinguishing between different classes; and (2) computing similarity between feature vectors to retrieve relevant videos. Unlike recognition algorithms, retrieval does not require specific action classification but focuses on finding similar videos. Consequently, retrieval algorithms prioritize efficiency and scale robustness—the primary challenges facing video retrieval systems.

The proposed method uses gradient features from spatio-temporal interest points to construct video feature vectors, introduces fuzzy clustering to build clustered feature vectors for improved robustness, and constructs a fuzzy weight matrix to modify Euclidean distance measures for more robust similarity computation. The workflow is shown in [Figure 2: see original paper].

2.1 Spatio-Temporal Interest Point Detection

We employ the commonly used 3D Harris operator to detect spatio-temporal interest points. The 3D Harris operator extends the 2D spatial Harris operator by adding a temporal dimension, sharing the same detection principles and implementation steps. First, scale-space representation is applied to the video:

$$L(\cdot;\sigma,\tau) = G(\cdot;\sigma,\tau) * I$$

where σ and τ are spatial and temporal scale parameters, respectively; $*$ denotes convolution; I is the input video; and G is the Gaussian kernel:

$$G(x,y,t;\sigma,\tau) = \frac{1}{\sqrt{(2\pi)^3\sigma^4\tau^2}} \exp\left(-\frac{x^2+y^2}{2\sigma^2} - \frac{t^2}{2\tau^2}\right)$$

The second-moment matrix M is then computed by expanding L using Taylor series:

$$M = G(\cdot;\sigma,\tau) * \begin{bmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2 \end{bmatrix}$$

The 3D Harris corner detector is obtained as:

$$R = \det(M) - k \cdot \mathrm{trace}^3(M)$$

where $\det(M)$ is the determinant, $\mathrm{trace}(M)$ is the trace, k is a constant (typically 0.04–0.06), and λ_1, λ_2, λ_3 are eigenvalues of M. Implementation details are provided in reference [10].

2.2 Feature Vector Extraction

We construct video feature descriptors using gradient information from spatio-temporal interest points. Upon detecting an interest point, a spatio-temporal cube is applied, and gradient descriptors are computed for each cube to determine the primary motion direction and scale. Results along each axis are concatenated to form a gradient vector for the interest point. Implementation follows reference [11]. Notably, we avoid dimensionality reduction to prevent feature loss. Instead, computational complexity is reduced through clustering during similarity computation, achieving efficiency comparable to dimensionality reduction while preserving feature integrity.

For any video v_i, let n denote the number of detected spatio-temporal interest points. The corresponding feature vector is $F_i = \{f_{i,j} | j = 1, 2, \dots, n\}$, where $f_{i,j}$ represents the feature vector extracted from the j-th interest point.

2.3 Similarity Computation

For each video' s feature vector, we first apply the fuzzy k-means clustering algorithm to the gradient vectors of spatio-temporal interest points, enhancing robustness (see reference [12]). With c cluster centers, the clustered feature vector becomes $F_i^c = \{f_{i,j}^c | j = 1, 2, \dots, c\}$, where $f_{i,j}^c$ denotes the j-th class of gradient vectors after clustering.

After clustering, we compute similarity between the query video' s clustered feature vector and those of database videos, sorting by similarity. Greater similarity indicates stronger correlation between videos. Typically, similarity is reflected by distance between feature vectors—shorter distances imply greater similarity. We adopt this principle for our similarity calculation.

Given that clustered feature vectors contain c classes of gradient vectors, and that variations in scale, position, and timing across different video captures prevent one-to-one correspondence between the c classes, we first compute distances between all pairs of gradient vector classes from two clustered feature vectors. The pair with minimum distance is selected to establish correspondence between the c classes. Additionally, considering that factors like scale and position may cause significant differences in gradient magnitudes (e.g., closer objects yield larger gradient magnitudes), we introduce fuzzy weights to normalize gradient vectors before computing normalized distances.

Let v_i denote the query video and v_j a database video, with clustered feature vectors F_i^c and F_j^c, respectively. The distance between the p-th gradient vector class of F_i^c and the q-th class of F_j^c is:

$$d_{v_i,v_j}(p,q) = \|f_{i,p}^c - f_{j,q}^c\|_2$$

The fuzzy weight for this pair is:

$$w_{v_i,v_j}(p,q) = \frac{1}{\left(d_{v_i,v_j}(p,q)\right)^f + \varepsilon}$$

where f is a fuzzy constant (set to 0.3 in this work) and ε is a small positive number (0.00001) to avoid division by zero.

A $c \times c$ fuzzy weight matrix W_{v_i,v_j} is constructed from all pairs. The matching index for the p-th gradient vector class in F_i^c is:

$$C(p) = \arg\min_{q \in C} w_{v_i,v_j}(p,q) \cdot d_{v_i,v_j}(p,q)$$

where C represents the set of available indices. Once a gradient vector class is matched, it is removed from subsequent matching iterations, ensuring each class is paired exactly once.

The overall distance between clustered feature vectors F_i^c and F_j^c is the average fuzzy distance across the c matched pairs:

$$d_{v_i,v_j} = \frac{1}{c}\sum_{k=1}^{c} w_{v_i,v_j}(k, C(k)) \cdot d_{v_i,v_j}(k, C(k))$$

Finally, similarity is inversely related to distance:

$$s_{v_i,v_j} = \frac{1}{d_{v_i,v_j} + \varepsilon}$$

2.4 Relevant Video Output

For a query video, we compute similarities with all database videos and rank them in descending order. The query margin parameter U specifies the number of retrieved videos. In this work, the top U videos with highest similarity are returned as retrieval results.

3 Experimental Analysis

As no dedicated benchmark database exists for human activity video retrieval, we conduct experiments on the widely used KTH database for activity recognition. This large-scale database contains 2,391 video clips, each showing a single person performing one of six actions: walking, jogging, running, boxing, hand-waving, and hand-clapping, performed by 25 subjects across four scenarios with variations in clothing and scale. Videos are 160×120 resolution at 25 fps.

We evaluate retrieval performance using precision, recall, and retrieval time—standard metrics in image retrieval. For each action class, the first 100 videos

serve as queries while the remainder form the retrieval database. The query margin parameter U is set to 20.

3.1 Retrieval Accuracy Testing

The proposed method includes a parameter c (number of cluster centers) that significantly impacts retrieval accuracy. Generally, larger c yields finer feature partitioning, improving between-class discrimination but reducing within-class tolerance. We determine the optimal c by comparing average precision and recall across different values. As shown in [Figure 4: see original paper], both metrics peak when $c = 6$, which we adopt as the optimal parameter. At this setting, the method achieves 78.1% average precision and 73.8% average recall.

3.2 Performance Comparison

Human activity video retrieval is a relatively recent research area with limited existing work. We compare our method against three representative approaches [5–7] under identical experimental conditions (database, query videos, $U = 20$, Intel i5 CPU, 16 GB RAM, Windows 7 64-bit, Visual Studio 2013). Results are summarized in , where "Average" represents the mean of precision and recall, and retrieval time is measured in seconds.

Method	Avg. Precision	Avg. Recall	Average	Avg. Retrieval Time
[5]	67.3%	59.4%	63.4%	–
[6]	71.6%	72.9%	72.3%	–
[7]	78.8%	69.7%	74.3%	–
Ours	78.1%	73.8%	76.0%	–

Our method achieves the highest average recall and, despite slightly lower precision than [7], delivers the highest combined average metric. Moreover, our method' s retrieval time is significantly lower than the compared methods, demonstrating substantially higher efficiency—a critical advantage for video retrieval systems.

References

[1] Liang J J, Xiong Y J, Yu D H. Research on ontology-based video retrieval technology [J]. Computer Engineering & Science, 2015, 37(10): 1940–1946.

[2] Chaaraoui A A, Climent-Pérez P, Flórez-Revuelta F. Silhouette-based human action recognition using sequences of key poses [J]. Pattern Recognition Letters, 2013, 34(15): 1799–1807.

[3] Chen C, Jafari R, Kehtarnavaz N. Improving human action recognition using fusion of depth camera and inertial sensors [J]. IEEE Trans on Human-Machine Systems, 2015, 45(1): 51–61.

[4] Chaaraoui A A. Evolutionary joint selection to improve human action recognition with RGB-D devices [J]. Expert Systems with Applications, 2014, 41(3): 786-794.

[5] Tang J, Shao L, Zhen X. Human action retrieval via efficient feature matching [C]// Proc of IEEE International Conference on Advanced Video and Signal Based Surveillance. 2013: 306-311.

[6] Jones S, Shao L, Du K. Active learning for human action retrieval using query pool selection [J]. Neurocomputing, 2014, 124(2): 89-96.

[7] Jiang Y G, Li Z, Chang S F. Modeling scene and object contexts for human action retrieval with few examples [J]. IEEE Trans on Circuits & Systems for Video Technology, 2011, 21(5): 674-681.

[8] Li R F, Wang L L, Wang K. Survey on human action recognition [J]. Pattern Recognition and Artificial Intelligence, 2014, 27(1): 35-48.

[9] Feng J G, Xiao J. Viewpoint-independent action recognition: a survey [J]. Journal of Image and Graphics, 2013, 18(2): 157-168.

[10] Sipiran I, Bustos B. Harris 3D: a robust extension of the Harris operator for interest point detection on 3D meshes [J]. The Visual Computer, 2011, 27(11): 963-976.

[11] Fang X, Tian Y, Wang Y, et al. Pair-wise event detection using cubic features and sequence discriminant learning [C]// Proc of IEEE International Conference on Multimedia and Expo. 2013: 1-6.

[12] Cebeci Z, Yildiz F. Comparison of K-means and fuzzy C-means algorithms on different cluster structures [J]. Journal of Agricultural Informatics, 2015, 6(3): 13-23.

Note: Figure translations are in progress. See original paper for figures.