키프레임의 딥특징을 사용하여 영화장면에서의 동작 인식

Action Recognition in Movie Scenes using Deep Features of Keyframes

울라 아민, 아마드 자밀, 무하마드 칸, 메흐무르 이르판*, 이미영, 박준렬, 백성욱*[1)]
Amin Ullah, Jamil Ahmad, Khan Muhammad, Irfan Mehmood,
Mi Young Lee, Jun Ryeol Park, Sung Wook Baik

(05006) 서울특별시 광진구 능동로 209 세종대학교 전자정보공학대학
(05006) 서울특별시 광진구 능동로 209 세종대학교 소프트웨어융합대학*
{qamin3797, jamilahmad, khanmuhammad}@sju.ac.kr,
{irfan, miylee}@sejong.ac.kr, jrpark3797@gmail.com, sbaik@sejong.ac.kr

요 약

최근에 컴퓨터 비젼 연구자들은 비디오 클립에서 인간의 행동 인식에 초점을 맞추었고, 감시 및 스포츠와 같은 다양한 영역의 응용 프로그램에 사용했습니다. 이 논문에서는 키 프레임의 딥특징을 사용하여 영화 클립에서 인간의 행동을 인식했습니다. 첫째, k-mean 클러스터링은 액션 비디오로부터 대표적인 프레임(키프레임)을 획득하는데 사용됩니다. 클러스터링은 액션 비디오로부터 중복된 프레임을 제거함으로써 알고리즘의 복잡성을 줄여줍니다. 둘째, 12편의 영화를 위해 Alexnet이라 불리는 컨볼루션 신경네트워크(CNN)모델을 미세 조정했습니다. 마지막으로 우리는 대표적인 액션 프레임을 CNN분류기에 제공했습니다. 이렇게 제안된 알고리즘은 Holywood2 데이터 세트에 대해 실험적으로 테스트 되었으며, 획득된 결과는 최첨단 기술로 만들어진 특징 추출 기반의 동작 인식방법과 비교하여 정확성 측면에서 더 우수하였습니다.

Abstract

Recently, researchers of computer vision have focused on human action recognition in video clips and have used it for applications in various domains such as surveillance and sports. In this paper, we have recognized human action in movies clips using deep features of keyframes. Firstly, k-mean clustering is used to achieve representative frames (keyframes) from action videos. Clustering removes redundant frames from action videos, thereby reducing the computational complexity of the algorithm. Secondly, we fine-tuned convolutional neural network (CNN) model called Alexnet for 12 movies actions. Finally, we feed those representative frames of action video clips to the fine-tuned CNN classifier. The proposed algorithm is experimentally tested on Holywood2 dataset

1) Corresponding author

논문투고일자 : 2017. 06. 09 심사일자 : 2017. 06. 12 게재확정일자 : 2017. 06. 26

and the obtained results are better in terms of accuracy compared to state-of-the-art hand crafted features extraction based action recognition methods.

키워드: 액션 인식, 딥 러닝, 대표 프레임, 미세조정된 CNN
Keyword: Action recognition, deep learning, representative frames, fine-tuned CNN

1. Introduction

Action recognition is a challenging research area under computer vision. Its purpose is recognition of a human action from video clip which is a sequence of images. Such system firstly analyzes the video clips to learn about the actions in the give clips and learn to identify similar actions. Automatic recognition of actions has led us to the development required application of modern world such as finding intruder in surveillance, categorization of video data, video indexing and retrieval, virtual coaches, understanding user environments, evaluation of robotic therapy as biofeedback device in dementia care, and the development of home assistant robots for ageing society [1]. Action recognition is a rapidly emerging area in computer vision. Vision based and pattern based recognition methods can be used for recognition of actions in videos. In the last decade, researchers have relay on identifying actions using simple features such as Spatio-temporal local features, Bag-of-Words (BOW) and motion based features. This only works for single action performed by one person in a scene, hence failing to deal with complex actions. This can be especially useful for surveillance of public locations such as subways, shopping centers, or parking lots in order to reduce crime, monitor traffic flow, and ensure security [2].

Aggarwal et al., [3] divided human actions into four levels for abstraction. They annotated actions as gestures, interactions, action, and group activities. Further, a gesture is defined as basic movement of parts of a person's body like moving an arm or raising their legs. An interaction is the action in which two persons are involved such as hugging someone or fighting each other. Such activities are achieved through a bunch of multiple peoples or objects such as group of people dancing or multiple cars passing on the road. We used a dataset of movies clips in which four type of actions are captured including stand up and sit down as gestures, hugging and kissing as interaction, and fighting related to others activities. In such scenarios, the system needs to be more complex for the recognition but the CNN classifier performs better because it extracts features and classify them at the same time, helping in classification of different type of categories at the same time.

2. Literature Review

In the early stages of action recognition, researchers have used hand crafted features along with many classifiers such as SVM, KNN, and decision tree. For instance, Klaser et al., [4] extended HoG to 3D and build the 3D HoG descriptor. It is based on histograms of 3D gradient orientations which is uniformly

quantized by regular polyhedrons in an integral video representation. Wang et al., [5] extracted the dense trajectories and motion boundary descriptors from the video as the repre-sentation. As the motion boundary descriptors can reduce the effects of camera motions effectively, it makes a huge progress in the realistic videos based action recognition and can be treated as the state-of-the-art. In [6], Igor et al., achieved action recognition using covariance of shape and motion with low dimensional features. They also have introduced a runtime optimizer for the real time execution using nearest neighbor classification. Lulu et al., [7] made a dataset using Kinect sensors for more than two thousand videos clips. They have fused spatio-temporal local features with Bag-of-Words (BOW) for dynamic time alignment of action recognition in videos. Laptev et at., [8] presented a technique for video classification and action recognition which is based on analyzing different recent methods including local space-time features, space-time pyramids and multichannel non-linear SVMs. This method achieved 91.8% accuracy on KTH dataset.

CNN models learn visual features from data by using back propagation neural network approach, which works for large data in an accurate way to correct parameter of the classifier. However we have some pre trained CNN models which are trained on large scale datasets. We can use those parameters as transform learning for our small datasets [9]. Those parameters promisingly extract visual features from new data very accurately. In [10], SFA learning was combined with three dimensional CNN for automated action representation and recognition. This method achieved state-of-the-art results on three public datasets including KTH, and UCF sports. Other types of supervised models include recurrent neural networks (RNNs). For instance, a method using RNN was proposed in [11] for skeleton-based action recognition. The human skeleton was divided into five parts and then separately fed into five subnets. The results of these subnets were fused into the higher layers and final representation was fed into the single layer.

3. Proposed Methodology

The proposed technique for action recognition in movie videos is divided into two main steps. Firstly, we extract representative frames from video by applying K-Mean [12] clustering algorithm. Secondly, we pass those represen-tative frames to a fine tuned Alexnet [13] CNN model which labels each frame. Finally, we classify video based on the maximum number of labels assign by CNN classifier. The framework for training proposed classifier is shown in Fig 1. The testing mechanism is illustrated in Fig 2.

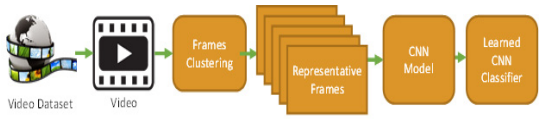


Fig 1. Training Framework

The proposed method has two phases: Firstly, we arrange all training data by finding its representative frames for each clip in one folder along with their labels. Then a fine-tuned Alexnet [13] is used to train the

classifier for this data. Fig 1 shows the method of training in the proposed technique. As a result of training phase, a trained classifier is obtained which is used for recognizing actions in new video clips. In testing phase, we get frames from video clips and cluster them for obtaining representative frames, which are passed to fine-tuned Alexnet model. Fig 2 shows the procedure of testing video clips for action recognition. CNN model provides prediction for each frame, where prediction of each frame is combined for final decision and assign maximum predicted labels as predicted action.

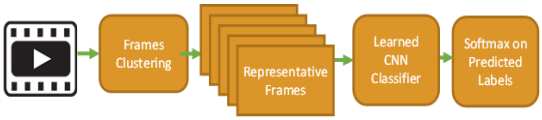


Fig 2. Testing Framework

3.1 Clustering

While working with video processing it is a challenging task to handle the complexity of the algorithm, due to the high resolution and more than 40 frame per sec. To handle this problem we have used clustering technique to first identify the representative frame and then just analyze those frame for the final classification of the action. This give us a low complexity for the per frame computation. We have extracted CNN features from frame in video with 10 frames jump in sec and used K-mean algorithm for clustering those frame. After we get clusters from video we calculate the Euclidean distance of frame of each cluster with its cluster's centroid, and select a frame having less distance with cluster centroid as a representative frame. Fig show the concept of clustering frame in Video Clip.

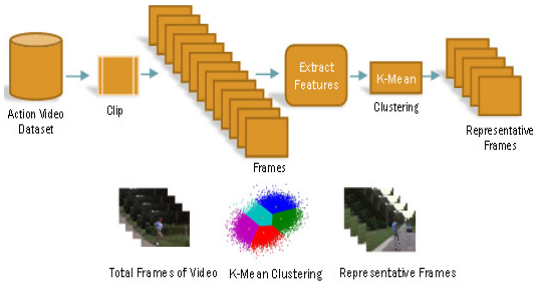


Fig 3. Getting representative frames using CNN features and K-means clustering

3.2 Fine tune AlexNet

The novel Alex model was trained for the classification of thousand classes using the ImageNet dataset of 1.28 million images. In the proposed method, we have used the concept of transform learning and fine-tuned AlexNet model for action recognition domain by changing some basic parameters and dataset [14]. We have used Hollywood2 dataset for training the model. AlexNet has five convolutions and two fully connected layers followed by a softmax layer, where layer 1 and 2 have conv pooling and norm, 3 and 4 have just conv and layer 5 has both conv and pooling. We have feed 128x128 images instead of model default images size which is 227x227, because the cropped frames are low resolution. With resizing of frames, we lost a lot of information in frames. Features are learnt on each layer by apply ReLU nonlinear activation function. From Fig 4, it can be seen that features are extracted from each region of the action frame. Some parts of action are identified in first layer and others on next layers. Indeed, a small portion is not hidden from the deep learning model. Finally we have changed dimension of softmax layer to the number of classes in Hollywood dataset.

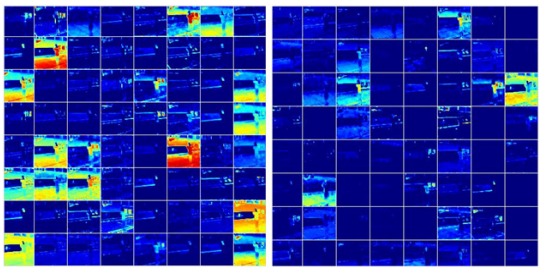


Fig 4. Feature maps for action clip frames

4. Experimental Evaluation

In this section, we have assessed the proposed technique by exploitation of confusion matrix. We have used a benchmark dataset of action recognition called Holywood2 [8] video dataset. We have randomly selected 50 to 60 test video clips from each category for the evaluation of proposed technique. The performance of the other state-of-the-art techniques has been compared with our proposed method of action recognition.

4.1 Hollywood2 Dataset

The proposed technique for action recognition is evaluated on Holywood2 [8] action videos dataset, which consists of 12 categories including car driving, eating, fighting, get out form car, hand shaking, hugging, phone call, sit up, sit down, and stand up. Each class has two to three hundred videos for training and one hundred for testing. The dataset is very challenging due to the overlapping visual contents in each category. For instance, an action of eating is performed in the same room where the action of hugging is captured. We have selected fifty to sixty test video clips from each class and discard some of the confusing video clips due to their difficulty in distinguishing the clear difference in their visual contents. For example, a man standing from the eating table is given the label "sit up" while the same man sitting in front of table is labeled as "eating". Such clips are not used in our training phase. The test video clips have been evaluated using confusion matrix for true positive, true negative, false positive, and false negative scores. The confusion matrix is given in Fig 5.

Fig 5 shows the overall performance of the proposed method, where we obtained 81.7%

Fig 5. Confusion matrix for proposed method.

	Car Drive	Eating	Fight	Get out Car	Hand Shake	Hug Person	Kiss	Phone Ans	Run	Sit Up	Sit Down	Stand Up	Truth Overall	Accuracy Precision
Car Drive	30	3	2	2	2	0	4	2	1	2	2	1	51	58.8%
Eating	2	39	2	0	3	0	4	1	2	4	2	1	60	65%
Fight	0	0	41	0	0	0	3	1	0	1	0	0	46	89.13%
Get out Car	3	2	1	43	0	4	0	2	0	5	0	2	62	69.35%
Hand hake	1	0	2	0	28	1	0	4	0	2	0	3	41	68.29%
Hug Person	0	0	0	0	0	45	0	0	3	2	1	1	52	86.53%
Kiss	0	0	0	0	0	3	50	2	0	0	2	2	57	87.71%
Phone Ans	2	5	1	3	1	1	0	43	1	0	1	0	58	74.13%
Run	0	3	1	2	1	0	0	0	37	0	0	0	44	84.09%
Sit Up	2	3	3	1	0	1	0	0	0	26	0	0	35	74.28%
Sit Down	0	0	0	0	2	1	0	0	0	0	37	0	43	86.04%
Stand Up	0	0	1	1	3	1	0	0	0	0	0	47	54	87.03%
Truth Overall	40	55	54	56	40	56	61	55	44	42	45	55	603	81.7%

total average accuracy for all classes. In Fig 5, the column "classification overall" shows the total number of test clips for each class. The diagonal shows the correct recognition of the corresponding action. In the experiments, we got less accuracy for class "eating table" which is 58.82%. This is due to the similarity in visual contents which match with other classes such as hugging a person or sitting down for eating on table, and sitting up from eating table. We got maximum accuracy for the class "running person" because it is very different from other classes. On the basis of overall accuracy, the proposed method is a good descriptor for action recognition in small video clips.

Table. 1 Performance comparisons with handcrafted features based action recognition methods

Actions \ Methods	HoG	HoG+ BoW	Proposed
Answer Phone	42.22%	55.21%	58.82%
Get Out Car	50.64%	38.1%	69%
Hand shaking	49.7%	49.62%	68%
Hug person	40.55%	59%	69%
Sit Down	55.45%	64%	68%
Sit up	39.44%	47.21%	74.28%
Stand up	50.88%	39.66%	87.03%
Car Drive	—	—	58.82%
Eating Table	—	—	65%
Fighting	—	—	86%
Running	—	—	84%
Kissing	—	—	87%

Table. 1 shows comparison with two handcrafted features extraction based action recognition methods. The recognition scores are taken from the reference paper [8] and are compared with our method. The accuracy for some categories is not reported, thus, we have put dashes in those columns. Histogram of oriented gradients (HoG) has low accuracy as compared to HoG with BoW. There is variation in accuracy for each category in other competing methods, however, the proposed method achieves better accuracy for all categories with less variation. Fig 6 shows some of the screenshots captured in testing of videos. The labels are given to each action after analyzing the representative frames of the underlying video clip.

Fig 6. Screenshots of different videos during evaluation.

5. Conclusion

In this paper, we proposed an action recognition system using CNN classifier by customizing Alexnet using transfer learning strategy. K-mean clustering algorithm and CNN features are intelligently used to remove redundancy of frames and achieve efficiency in computation [15]. We used Holywood2 movie clips dataset for training the proposed classifier and evaluation. The obtained results show that proposed technique is effective for action recognition when compared with some existing methods. The efficiency and accuracy can be further improved by intelligent fine-tuning of light-weighted CNN models such as Squeeze-

Net along with additional time space information of CNN features.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIP) (No. 2016R1A2B4011712).

References

〔1〕 Ramanathan, M., W.-Y. Yau, and E.K. Teoh, Human action recognition with video data: research and evaluation challenges. IEEE Transactions on Human-Machine Systems, 2014. 44(5): p. 650-663.

〔2〕 Youssef, M. and V. Asari, Human action recognition using hull convexity defect features with multi-modality setups. Pattern Recognition Letters, 2013. 34(15): p. 1971-1979.

〔3〕 Aggarwal, J.K. and Q. Cai. Human motion analysis: A review. in Nonrigid and Articulated Motion Workshop, 1997. Proceedings., IEEE. 1997. IEEE.

〔4〕 Klaser, A., M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. in BMVC 2008-19th British Machine Vision Conference. 2008. British Machine Vision Association

〔5〕 Wang, H., et al., Dense trajectories and motion boundary descriptors for action recognition. International journal of computer vision, 2013. 103(1): p. 60-79.

〔6〕 Kviatkovsky, I., E. Rivlin, and I. Shimshoni, Online action recognition using covariance of shape and motion. Computer Vision and Image Understanding, 2014. 129: p. 15-26.

〔7〕 Chen, L., H. Wei, and J. Ferryman, ReadingAct RGB-D action dataset and human action recognition from local features. Pattern Recognition Letters, 2014. 50: p. 159-169.

〔8〕 Laptev, I., et al. Learning realistic human actions from movies. in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. 2008. IEEE.

〔9〕 Amin ullah, M.S. Content based Image Retrieval using Local Shape and Color Features in Lab* Color Space. in The 2nd International Conference on Next Generation Computing 2017. 2017. Korean Institute of Next Generation Computing, Korea.

〔10〕 Sun, L., et al. DL-SFA: deeply-learned slow feature analysis for action recognition. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.

〔11〕 Du, Y., W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

〔12〕 Hartigan, J.A. and M.A. Wong, Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 1979. 28(1): p. 100-108.

〔13〕 Krizhevsky, A., I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. in Advances in neural information processing systems. 2012.

〔14〕 Amin Ullah, N.R., Jamil Ahmad, Mi Young Lee, Sung Wook Baik, Analyzing Pedestrian Parts using Deep Features for Person

Re-Identification. Korean Institute of Next Generation Computing, 2017. 13(2): p. 10.
[15] Ejaz, N., I. Mehmood, and S.W. Baik, Efficient visual attention based framework for extracting key frames from videos. Signal Processing: Image Communication, 2013. 28(1): p. 34-44

저자소개

◆ Amin Ullah(울라 아민)

• He received his BS degree in Computer Science from Islamia College, Peshawar in 2016. Currently, he is pursuing his MS leading to PhD degree from Intelligent Media Lab, Sejong University, South Korea. His research interests include content-based image retrieval, deep learning, and computer vision.

◆ Jamil Ahmad(아마드 자밀)

• He received his BCS degree in Computer Science from the University of Peshawar, Pakistan in 2008. He received his Master's degree in Computer Science with specialization in image processing from Islamia College, Peshawar, Pakistan. Currently, he is pursuing PhD degree in digital contents from Sejong University, Seoul, Korea. His research interests include image analysis, semantic image representation, deep learning, and content based multimedia retrieval.

◆ Khan Muhammad(무하마드 칸)

• He received his BS degree in Computer Science from Islamia College Peshawar, Pakistan. He is currently pursuing MS leading to PhD degree in digital contents from in Sejong University, Seoul, Korea. His research interests include digital image and video processing, Information Hiding and Security.

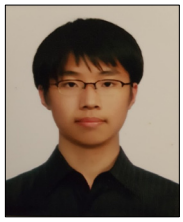
◆ Irfan Mehmood(메흐무르 이르판)

• He received his BS degree in Computer Science from National University of Computer and Emerging Sciences, Pakistan. He completed Ph.D. degree from Sejong University, Seoul, Korea. Dr. Irfan is Assistant Professor in College of Electronics and Information Engineering at Sejong University, Seoul, South Korea. His research interests include video and medical image processing, big data analysis, and visual information summarization.

◆ Mi Young Lee

• She is a research professor at Sejong University. She received her MS and PhD degree in the Image and Information Engineering at Pusan National University. Her research interests include Interactive Contents, deep learning, and computer vision, UI, UX and Developing Digital Contents.

◆ Jun Ryeol Park



• He is pursuing bachelor's degree in digital contents from Sejong University, Seoul, Korea. His reasearch interests include image analysis, machine learning, metaheuristic.

◆ Sung Wook Baik

• He is a PhD in Information Technology and Engineering from George Mason University. He is a professor in the College of Sofrware and Conver-gence Technology at Sejong University. His research interests include Computer vision, Pattern recognition, Computer game and AI.