International Journal of Engineering

Journal Homepage: www.ije.ir

Human Action Recognition using Prominent Camera

P. S. Kavimandan*[a], R. Kapoor[b], K. Yadav[a]

[a] *Indira Gandhi Delhi Technical University for Women, Delhi, India*
[b] *Delhi Technological University, Delhi, India*

PAPER INFO

Paper history:
Received 19 August 2020
Received in revised form 03 November 2020
Accepted 05 November 2020

Keywords:
Action Recognitio
Modified Bag-of-Words
Prominent camer
Support Vector Machine

ABSTRACT

Human action recognition has undoubtedly been under research for a long time. The reason being its vast applications such as visual surveillance, security, video retrieval, human interaction with machine/robot in the entertainment sector, content-based video compression, and many more. Multiple cameras are used to overcome human action recognition challenges such as occlusion and variation in viewpoint. The use of multiple cameras overloads the system with a large amount of data, thus a good recognition rate is achieved with cost (in terms of both computation and data) as the overhead. In this research, we propose a methodology to improve the action recognition rate by using a single camera from multiple camera environments. We applied a modified bag-of-visual-words based action recognition method with the Radial Basis Function-Support Vector Machine (RBF-SVM) as a classifier. Our experiment on a standard and publicly available dataset with multiple cameras shows an improved recognition rate compared to other state-of-the-art methods.

***doi**: 10. 5829/ije.2021.34.02b.14*

1. INTRODUCTION

The development of technology has made the camera a very easily available gadget, to the point that almost everybody these days uses a digital camera for capturing pictures or recording videos on a daily basis. In profession video recording for film production many cameras are in use simultaneously. The director of the film decides which camera has captured the best scene for a particular moment. But it is not always possible for a human administrator to be present for making such decisions. Thus, an automated machine that can decide which camera has produced the best scene can be valuable. The best information depends on the best view which in turn depends on several points. A view from one angle for a particular action may not be a good view from another angle for that same action. Thus the analysis of the video is required to select the best camera for the particular action. With the advent of technology and the rising number of videos, analyzing a video for the purpose of action recognition has gained tremendous importance in the field of computer vision. Action recognition is quite helpful especially in fields such as surveillance, security, robotics, etc. It also plays a vital role in intelligent systems. The aim here is to identify the category of actions performed in the video from the viewpoint of a selected camera.

In this work, we propose a methodology to recognize human actions performed by an actor by selecting a prominent camera in the multi-camera scenario. This technique helps to avoid the data processing captured by all the cameras. We rely here on the principle that different cameras capture different views, all of which are not good enough to recognize the action. Thus, we develop a score for each of the cameras based on two factors: limb visibility and limb movement. Depending on the score, one of the cameras is chosen as the 'prominent camera and selected for further processing. To testify the methodology, standard databases such as IXMAS [23] are used. The structure of the remaining paper is as follows: literature survey is discussed in Section 2; Section 3 explains the methodology in detail; in Section 4 results are discussed; followed by a conclusion and future work in Section 5.

*Corresponding Author Email: *pranoti.sk@gmail.com* (P. S. Kavimandan)

Please cite this article as: P. S. Kavimandan, R. Kapoor, K. Yadav, Human Action Recognition using Prominent Camera, International Journal of Engineering, B: Applications Vol. 34, No. 02, (2021) 427-432

2. RELATED WORKS

Action recognition from multiple views has been broadly studied in two approaches, the 2D approach and the 3D approach. 2D approaches deal with methods that use data from multiple cameras captured independently; whereas 3D approaches deal with multiple cameras that have been set up at a fixed location [1]. 2D approaches can be divided into two categories. First, the view independent category where data from all cameras is captured independently. And then either action is represented with view-invariant features [2-5] or several classifiers are fused for classification [6]. In Yilmaz and Shah [2], a moving camera is used for capturing data. Thus along with humans, the camera trajectory is also moving. Authors proposed the geometry of dynamic scenes to recognize human action in a number of challenging sequences. Ashraf et al. [3] have used the notion of projective depth to implement the view-invariant action recognition method. Projective depth is unaffected by the camera's orientation and thus has been used to recognize similar actions. In Junejo et al. [4], self similarity descriptors are used since, according to the authors, they are very stable and do not require a correlation between multiple views. A novel action descriptor is constructed using a temporal Laplacian Eigen map that converts view-dependent videos to a stylistic invariant embedded manifold for every single view in Lewandowski et al. [5]. Ahmad and Lee [6] proposed a novel method to recognize human action from a random view by combining features from silhouettes and optical flow.

Other approaches used universal classifiers to recognize actions from the data received from each camera [7-10]. A novel algorithm based kernelized structural SVM is used as a classifier to recognize human action from a random view in Wu and Jia [7]. Zhu et al. [8] proposed a novel multi-sensor fusion method and the universal classifier random forest is used for action recognition. Action videos are represented as prototypes of human body postures using self-organizing maps that are spatially related in Iosifidis et al. [9]. Subsequently, action classification is done using multi-layer perceptrons in a Neural Network. Wang et al. [10] used cross-view action recognition with a K-NN-like classifier. Local features extracted from the input video form a bag of word using k-means clustering. The action is recognised bases on the transfer probability between visual words.

3D approaches usually combine all the visual information (2D human poses) gathered from each camera to represent it in the form of features that are eventually used for action recognition. Thus actions are represented as consecutive 2D human poses. Volumetric data helps to generate a system that is robust as proposed by Pierobon et al. [11]. It also helps to overcome the problem of self-occlusion that is obvious in the multi-camera environment. They have used only posture dependent characteristics as descriptors. Weinland et al. [12] have introduced Motion History Volumes (MHV) as a descriptor that is viewpoint independent. They used Fourier transforms in cylindrical co-ordinates around a vertical axis to align and compare their results. Another representation such as spherical harmonics has been used as a descriptor by Kazhdan et al. [13] with the purpose of avoiding calculations for the optimal alignment which are unfeasible. Their descriptor is rotation invariant. Gkalelis et al. [14] have used fuzzy vector quantization (FVQ) and linear discriminant analysis (LDA) to model and recognize different human movements. They have exploited rich data in multi-view videos and have used Discrete Fourier Transform (DFT), since it is circular and shift-invariant, to solve correspondence issues between training and testing samples. Holte et al. [15] took advantage of intensity and depth maps both captured by SwissRanger SR4000 camera. According to them this combination and the use of Motion Context (HMC) as an action descriptor improved the detection quality. Holte et al. [16] again used HMC along with 3D Motion Context (3D-MC) as the motion descriptor. Few authors, Feizi [17] and Sezavar et al. [18] have implemented Convolution Neural Network (CNN) for their methodologies, but Support Vector Machine (SVM) seems to be the better choice since the main focus is to reduce the computation time and cost [19]. Approaches based on 3D methods are found to be superior to approach based on 2D methods concerning recognition accuracy [20]. To deal with the important challenges in human action recognition such as variation in viewpoint and occlusion, the basic solution is to use multiple cameras, and thus literature concerning multiple cameras has been studied. But multiple camera usage hampers the complexity and running time of the system. To deal with it, in this research paper, we propose a human recognition technique to chose a prominent camera from the multi camera environment. We propose to gather initial data from multiple cameras but to process the data only from a prominent camera.

3. METHODOLOGY

Figure 1 shows the block diagram of the proposed methodology. The video along with the action expected

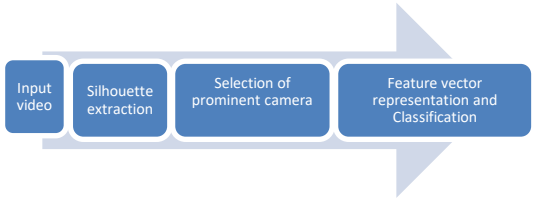

Figure 1. Flowchart of the proposed methodology

to be recognized is given as input. From the action, video silhouettes are extracted from all of the available cameras. Based on extracted silhouettes, information regarding limb visibility and movements can be discovered. We later calculate two scores namely, temporal score and spatial score for each camera. Based on the combination of these scores, a total score for each camera is calculated. The camera with the highest score is chosen as the most prominent camera, and the feature vector is represented for that prominent camera. These feature vectors act as input to the RBF-SVM classifier to recognize human action.

3. 1. Silhouettes Extraction The initial step in the methodology is silhouette extraction. This step requires background subtraction. Background subtraction has its own set of challenges such as variations in lighting, noise, etc. In the proposed methodology, the background has been modeled using a very popular technique known as the Gaussian Mixture Model (GMM) [21] since it is considered to be the most trustworthy background estimation method. This GMM technique is an adaptive mixture model that helps to deal with the problem of lighting variations, motion repetition, etc.

3. 2. Selection of a Prominent Camera To select a prominent camera, first of all, a score for each camera is developed for every camera on two factors. One is the limb movement and the other is limb visibility. We rely on the fact that when limb movement is significant, we can say that the silhouettes in the video will also move significantly. Thus a good motion of limbs indicates that the camera capturing it has a good chance of recognizing the action successfully and in lesser time. Lesser time here implies that the data from the camera other than the prominent camera need not be processed. A camera will develop a good score if it captures significant limb movement. This is called a score for temporal measure (C_t). Since we are interested in the limb movements, we picked out cameras which views show significant variations in movement, be it anywhere in the camera view. To calculate the limb movement, we need to develop the Motion Energy Image (MEI) [22]. MEI is found out using the following Equation (1):

$$MEI(x,y) = \bigcup_{t=0}^{\tau-1} B(x,y,t) \quad (1)$$

where, *B(x, y, t)* is a sequence of a binary images that highlights the area where motion has occurred. Let the most prevalent frame concerning the number of pixels be denoted by $B_{max}(x,y)$, thus we define the confidence score of the camera for temporal measure C_t by the following Equation (2):

$$C_t = 1 - \frac{\sum_{x,y} B_{max}(x,y)}{\sum_{x,y} MEI(x,y)} \quad (2)$$

The other fact that we rely on is that when limb visibility is good, the silhouettes tend to generate a concave profile. To find out the concavity of the shape, we define a spatial measure called a confidence score for the spatial measure $C_s$as given in Equation (3):

$$C_s = 1 - \frac{Vol_{st}}{Vol_{ch}} \quad (3)$$

where Vol_{st} is the Spatio-temporal volume and Vol_{ch} is the convex hull. Both C_t and $C_s$scores are delimited by 1, to always keep them positive. Both of these confidence scores for each camera are multiplied to calculate Confidence Score for a camera (C_c). Thus Confidence Score for a camera (C_c) is a combination of the Confidence score for temporal measure (C_t) and the Confidence score for spatial measure (C_s) as shown in Equation (4):

$$C_c = C_t . C_s \quad (4)$$

Thus we get the final score of each camera. Based on this score, the camera with the highest score will be chosen as the 'prominent camera' and feature vectors from that particular camera are processed further rather than processing data from all cameras. To recognize an action, we are using the modified bag-of-visual-words method described in literature [23].

3. 3. Feature Vector Representation and Classification Bag-of-visual-words methods are popular but these methods fail to preserve the information related to the geometry of the structure. Thus the method described in literature [23] is used to overcome this drawback. According to this method, a Harris 3D detector has been used to extract spatiotemporal points of interest.

The Histogram of Oriented Gradients (HOG) and Histogram of Optical Flow (HOF) descriptors are used to express interest points. Thus a bag-of-visual-word is created using these spatiotemporal interest points for representing an action and the contribution of cluster points is calculated. Depending on the difference among them the contextual distance among the points of a cluster is determined. Directed graphs are then created which are described by Laplacian. A Radial Basis Function Support Vector Machine (RBF-SVM) is fed with the feature vector corresponding to those Laplacians for action recognition.

4. RESULT DISCUSSION

The dataset used for validation is IXMAS [24]. It consists data of 11 actions performed by 5 male and 5 female actors from five static cameras. Figure 2(a-e) shows frames from the IXMAS dataset, and the actor here, namely ‘Daniel’, is acting a 'kick'. In Figure 3(a-e) the same actor is acting a 'punch'. Frames of all five cameras

have been shown for both of the actions, kick and punch in the figures. Figure 4(a-d) shows the result of the computation of temporal and spatial confidence scores C_t and C_s for different cameras. Table 1 shows temporal and spatial scores C_t and C_s calculated for all the five cameras, Camera 0-Camera 4 of the IXMAS dataset for

(a) Camera 0 (b) Camera 1 (c) Camera 2 (d) Camera 3 (e) Camera 4

Figure 2. (a-e) Frames from IXMAS dataset with actor Daniel acting 'kick' in all 5 cameras

(a) Camera 0 (b) Camera 1 (c) Camera 2 (d) Camera 3 (e) Camera 4

Figure 3. (a-e) Frames from IXMAS dataset with actor Daniel acting 'punch' in all 5 cameras, (a) C_t = 0.62, (b) C_t = 0.48, (c) C_s = 0.58, (d) C_s = 0.50

(a) (b) (c) (d)

Figure 4. Figure showing scores for different cameras; (a) C_t score for Daniel's kick action in Camera 0 (Figure 2 a); (b) C_t score for Daniel's kick action in Camera 2 (Figure 2 c); (c) C_s score for Daniel's punch action in Camera 0 (Figure 3 a); (d) C_s score for Daniel's punch action in Camera 3 (Figure 3 d)

TABLE 1. Temporal and spatial scores C_t and C_s and the final confidence score C_c calculated for all 5 cameras Camera 0-Camera 4 of IXMAS dataset. Prominent camera's Confidence score C_c is highlighted.

Action	Camera 0			Camera 1			Camera 2			Camera 3			Camera 4		
	Ct	Cs	Cc	Ct	Cs	Cc	Ct	Cs	Cc	Ct	Cs	Cc	Ct	Cs	Cc
Check watch	0.38	0.22	0.083	0.55	0.42	0.231	0.49	0.50	0.245	0.60	0.45	0.270	0.29	0.45	0.130
Cross arms	0.30	0.27	0.081	0.50	0.53	0.265	0.54	0.55	0.297	0.59	0.50	0.295	0.42	0.66	0.277
Scratch head	0.33	0.40	0.132	0.52	0.32	0.166	0.56	0.51	0.285	0.55	0.61	0.335	0.45	0.45	0.202
Sit down	0.54	0.52	0.280	0.53	0.56	0.296	0.52	0.58	0.301	0.56	0.60	0.336	0.32	0.28	0.089
Get up	0.28	0.35	0.098	0.52	0.55	0.286	0.50	0.52	0.260	0.51	0.49	0.249	0.49	0.52	0.254
Turn Around	0.34	0.47	0.159	0.36	0.45	0.162	0.45	0.57	0.256	0.52	0.59	0.306	0.27	0.31	0.083
Walk	0.53	0.50	0.265	0.55	0.49	0.269	0.53	0.49	0.259	0.56	0.50	0.280	0.30	0.42	0.126
Wave	0.50	0.58	0.290	0.45	0.59	0.265	0.42	0.54	0.226	0.60	0.58	0.348	0.49	0.60	0.294
Punch	0.55	0.58	0.319	0.58	0.61	0.353	0.45	0.48	0.216	0.49	0.50	0.245	0.52	0.60	0.312
Kick	0.62	0.58	0.359	0.60	0.58	0.348	0.48	0.50	0.240	0.52	0.45	0.234	0.49	0.54	0.264

10 actions such as checking the watch, crossing arms, scratching head, sitting down, getting up, turning around, walking, waving, punching and kicking. Prominent camera's confidence score C_c is highlighted. Figure 5 shows the graphical results for C_t, C_s, and C_c of all the five cameras for 10 actions in the IXMAS dataset. It can be noted from the graph that no single camera can be labeled as a prominent camera. It depends on the action to be recognized as in which camera would be the prominent camera. Table 2 shows the comparison of the proposed method with the other methods. As can be seen from the table, Camera 0 gives a good accuracy of 90.8% while recognizing the action 'Kick'. Camera 1 is good for actions 'Get up' and 'Punch' whereas Camera 2 for 'Cross arms' with an accuracy of 90.6% and 92.4% respectively. For all other actions, Camera 3 gives a worthy accuracy of 91.2%. Thus, we can also observe that if the prominent camera is chosen for action recognition, significant computation can be avoided by not processing the data from cameras other than the prominent camera.

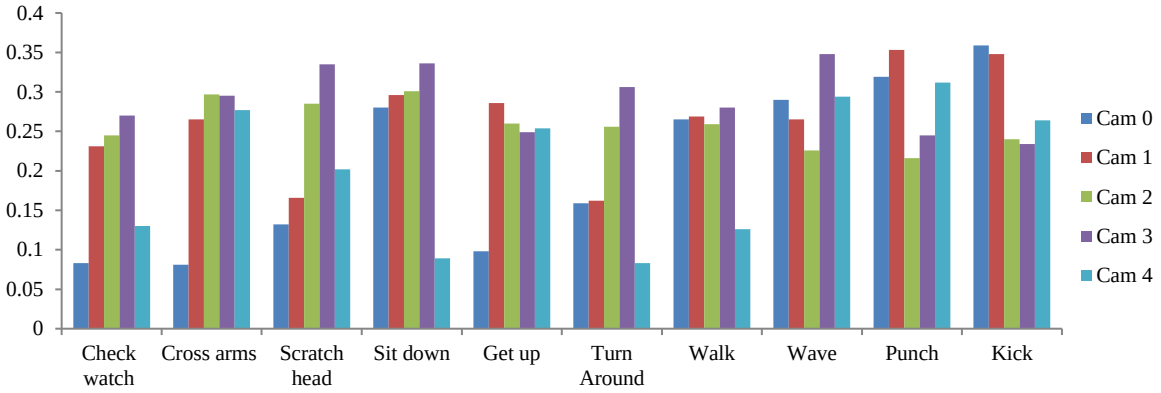

Figure 5. Shows the graphical results for confidence score C_c of all the five cameras for 10 actions in the IXMAS dataset

TABLE 2. Comparison of Proposed Method with other methods

Method	Accuracy (in Percentage)				
	Camera 0	**Camera 1**	**Camera 2**	**Camera 3**	**Camera 4**
Junejo I. et al. [4]	74.8	74.5	74.8	70.6	61.2
Wu, X. et al. [7]	86.5	83.8	86.1	84.5	87.4
Wang J. et al. [10]	88.4	85.3	88.3	86.5	87.2
Proposed Method	90.8 (Kick)	90.6 (Get Up, Punch)	92.4 (Cross Arms)	91.2 (All other actions)	90.6

5. CONCLUSION AND FUTURE WORK

In this paper, we have worked on the methodology of recognizing an action in an environment with multiple cameras. When multiple cameras are considered, along with their advantages comes the major disadvantage of computing time and cost because of the large amount of data. In our proposed methodology we have used the single camera with the most significant information for action recognition. We have given a score to each camera and based on this score the camera with the most significant information is chosen. As shown in Table 1, the prominent camera has been highlighted in yellow. The prominent camera may differ depending on the action, for example, the prominent camera in case of the action 'Checking watch' is Camera number 3 whereas, for the action 'Kich' it is Camera number 0. Thus, we processed the data from only a prominent camera among all the cameras to reduce the processing time and cost. In the future, we can fuse information from more than one prominent camera to further improve the recognition rate.

6. REFERENCES

1. Iosifidis, A., Anastasios T. and Ioannis P., "Multi-view Human Action Recognition: A Survey." 2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (2013), 522-525, doi: 10.1109/IIH-MSP.2013.135.

2. Yilmaz A. and Shah M., "Recognizing human actions in videos acquired by uncalibrated moving cameras" Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, Beijing, (2005), 150-157, doi: 10.1109/ICCV.2005.201.
3. Ashraf, N., Sun, C. and Hassan, F., "View Invariant Action Recognition Using Projective Depth" ***Computer Vision and Image Understanding***, Vol. 123, (2014) doi.:10.1016/j.cviu.2014.03.005.
4. Junejo, I., Dexter, E., Laptev, Ivan and Pérez, P., "View-Independent Action Recognition from Temporal Self-Similarities," In IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, No. 1, (2011), 172-185, doi.: 10.1109/TPAMI.2010.68.
5. Lewandowski, M., Dimitrios, M. and Jean-Christophe, N., "View and Style-Independent Action Manifolds for Human Activity Recognition." ECCV (2010). doi.org/10.1007/978-3-642-15567-3_40
6. Ahmad, M. and Lee, S. W., "HMM-based human action recognition using multiview image sequences", In Proceedings - 18th International Conference on Pattern Recognition, ICPR 2006, (2006), 263-266. doi.:10.1109/ICPR.2006.630
7. Wu, X. and Jia, Y., "View-Invariant Action Recognition Using Latent Kernelized Structural SVM", Fitzgibbon A., Lazebnik S., Perona P., Sato Y., Schmid C. (eds) Computer Vision – ECCV 2012, Lecture Notes in Computer Science, Vol. 7576. (2012) Springer, Berlin, Heidelberg. doi.: 10.1007/978-3-642-33715-4_30
8. Zhu, F., Shao, L. and Lin, M. "Multi-view action recognition Using local similarity random forests and sensor fusion", ***Pattern Recognition Letters***. Vol. 34, (2013) 20-24. doi.: 10.1016/j.patrec.2012.04.016.
9. Iosifidis, A., Tefas, A. and Pitas, I., "View-Invariant Action Recognition Based on Artificial Neural Networks," ***IEEE Transactions on Neural Networks and Learning System***s, Vol. 23, No. 3, (2012). 412-424, doi.: 10.1109/TNNLS.2011.2181865.
10. Wang, J., Zheng, H., Gao, J. and Cen, J., "Cross-View Recognition based ona Statistical Translation Framework", ***IEEE Transactions on Circuits and Systems for Video Technology***, Vol 26, (2014) doi. : 10.1109/TCSVT.2014.2382984
11. Pierobon, M. Marcon, M., Sarti, A. and Tubaro, S., "3-D Body Posture Tracking For Human Action Template Matching," 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Toulouse, (2006) doi.: 10.1109/ICASSP.2006.1660389.
12. Weinland, D., Ronfard, R. and Boyer, E., "Free viewpoint action recognition using motion history volumes", ***Computer Vision and Image Understanding***, Vol. 104, No. 2-3, (2006), 249-257, doi :https://doi.org/10.1016/j.cviu.2006.07.013
13. Kazhdan, M.M., Funkhouser, T.A., and Rusinkiewicz, S., "Rotation Invariant Spherical Harmonic Representation of 3D Shape Descriptors", Symposium on Geometry Processing, (2003) doi.:10.2312/SGP/SGP03/156-165
14. Gkalelis, N. Nokilaidis, N. and Pitas, I., "View independent human movement recognition from multi-view video exploiting a circular invariant posture representation", IEEE International Conference on Multimedia and Expo, (2009), 394-397. doi.: 10.1109/ICME.2009.5202517.
15. Holte M.B., Moeslund, T.B. and Fihl, P., "View-invariant gesture recognition using 3D optical flow and harmonic motion context", ***Computer Vision and Image Understanding***, Vol. 114, No. 12, (2010), 1353-136, https://doi.org/10.1016/j.cviu.2010.07.012
16. Holte, M.B., Moeslund, T., Nikolaidis, N. and Pitas, I., "3D Human Action Recognition for Multi-view Camera Systems", International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission, (2011), 342-349. doi.: 10.1109/3DIMPVT.2011.50.
17. Feizi, A., "Convolutional Gating Network for Object Tracking", ***International Journal of Engineering, Transactions A: Basics***, Vol. 32, No. 7, (2019), 931-939, https://dx.doi.org/10.5829/ije.2019.32.07a.05
18. Sezavar, A., Farsi, H. and Mohamadzadeh, S., "A Modified Grasshopper Optimization Algorithm Combined with CNN for Content Based Image Retrieval," ***International Journal of Engineering, Transactions A: Basics***, Vol. 32, No. 7, (2019), 924-930, https://dx.doi.org/10.5829/ije.2019.32.07a.04
19. Anding, K., Kuritcyn, P. and Garten, D., "Using artificial intelligence strategies for process-related automated inspection in the production environment", ***Journal of Physics: Conference Series***, Vol. 772, (2016). doi.:10.1088/1742-6596/772/1/012026.
20. Holte, M.B., Moeslund, T., Tran, C. and Trivedi M.M., "Human Action Recognition using Multiple Views: A Comparative Perspective on Recent Developments", MM'11-Proceedings of the 2011 ACM Multimedia Conference and Co-Located Workshops, (2011), doi.: 10.1145/2072572.2072588
21. Chen, Z. and Ellis, T.J., "A self-adaptive Gaussian mixture model", ***Computer Vision and Image Understanding***., (2014), Vol. 122, 35-46. https://doi.org/10.1016/j.cviu.2014.01.004
22. Bobick, A.F. and Davis, J.W., "The recognition of human movement using temporal templates," ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol. 23, No. 3, (2001), 257-267. doi.: 10.1109/34.910878.
23. Mishra, O., Kapoor, R. and Tripathi, M.M., "Human Action Recognition Using Modified Bag of Visual Word based on Spectral Perception", ***International Journal of Image, Graphics and Signal Processing***, Vol 11, (2019), 34-43. doi.: 10.5815/ijigsp.2019.09.04.
24. Weinland, D., Ronfard, R. and Boyer, E., "Free viewpoint action recognition using motion history volumes", ***Computer Vision and Image Understanding***, Vol. 104, No. 2-3, (2006), 249-257. doi: https://doi.org/10.1016/j.cviu.2006.07.013

Persian Abstract

چکیده

تشخیص عمل انسان بدون شک مدتهاست که تحت تحقیق بوده است. دلیل آن، کاربردهای گسترده آن مانند نظارت بصری، امنیت، بازیابی ویدئو، تعامل انسان با ماشین / ربات در بخش سرگرمی، فشرده‌سازی فیلم مبتنی بر محتوا و موارد دیگر است. از چندین دوربین برای غلبه بر چالش‌های تشخیص عملکرد انسان مانند انسداد و تغییر در دیدگاه استفاده می‌شود. استفاده از چندین دوربین سیستم را با مقدار زیادی داده بیش از حد بار می‌کند، بنابراین میزان شناخت خوبی با هزینه (از نظر محاسبه و داده) به عنوان سربار حاصل می‌شود. در این تحقیق، ما یک روش برای بهبود میزان تشخیص عملکرد با استفاده از یک دوربین منفرد از چندین محیط دوربین پیشنهاد می‌کنیم. ما یک روش اصلاح شده عمل مبتنی بر کلمات تصویری اصلاح شده با ماشین بردار عملکرد پشتیبانی-شعاعی (RBF-SVM) را به عنوان طبقه‌بندی اعمال کردیم. آزمایش ما بر روی یک مجموعه داده استاندارد و در دسترس عموم با چندین دوربین، میزان تشخیص بهتر در مقایسه با سایر روش‌های پیشرفته را نشان می‌دهد.