HUMAN ACTIVITY RECOGNITION USING STANDARD DESCRIPTORS OF MPEG CDVS

Łukasz Kamiński, Sławomir Maćkowiak, Marek Domański
Poznań University of Technology, Polanka 3, 60-965 Poznań Poland
{lkaminski, smack}@multimedia.edu.pl, marek.domanski@put.poznan.pl

ABSTRACT

In this paper, a novel human activity recognition method is presented. The proposed solution uses a single stationary camera in order to detect common human activities like: hand waving, walking, running etc. Unlike other methods which use different kinds of characteristic point descriptors in order to describe human poses, the proposed solution uses a CDVS descriptor which is part of the MPEG-7 standard. This allows the efficient calculation of a compact descriptor in a camera. The main goal of this work is to propose an efficient application of CDVS to describe and recognize different human activities. The experiments were performed on the Weizmann and KTH datasets. The obtained results prove the high accuracy of human activity recognition system with CDVS.

***Index Terms*—** activity recognition, CDVS, KTH dataset, Weizmann dataset

1. INTRODUCTION

In the areas of e-health technology and video surveillance, automated systems for observing motions of ill, elderly or handicapped people, pedestrian traffic and detecting dangerous actions are becoming very important these days. Many pedestrian areas are currently equipped with surveillance cameras. However, all the image understanding and risk detection is left to the security personnel. Similarly, in health surveillance, health and risk assessment is left to medical personnel. Such observations are very strenuous for human observers, as they require concentration over a long period of time. Moreover, human observers are very prone to mistakes that are extremely danger in the remote health surveillance. Therefore, it is a good reason to develop automated, intelligent, vision-based monitoring systems that can aid a human user in the process of risk detection and analysis.

The majority of current automated video surveillance systems can process video sequences and perform almost all key low-level functions, such as motion detection and segmentation, object tracking, and object classification with good accuracy. However, technical interest in video surveillance for both e-health and security, has shifted from such low-level functions to a more complex scene analysis, e.g., human action recognition.

Human activity recognition has been studied extensively over the last years. Human activity is related to the motion of the body. It is natural to use the methods which track the motion of characteristic points. However, these methods are computationally complex, and therefore very expensive. Based on the recent research in human activity recognition, the methods that do not use motion analysis can be roughly categorized into two main different categories based on the nature of the feature used for classification: texture, and shape, or a combination of both.

The Harris 3D detector [3], Hessian detectors [4], edge detector, corners detectors etc. represent local textural features. HOG/HOF (Histogram of Oriented Gradients/ Histogram of Optical Flow) [22] compute histograms of gradient and histograms of flow, 3D-SIFT (three dimensional Scale-Invariant Feature Transform) [5] extended the 2D-SIFT (two dimensional Scale-Invariant Feature Transform) descriptor from static images to video sequences, SURF (Speeded-Up Robust Features) descriptor [6] provides comparable or even better results than SIFT, while it can be calculated in a relatively efficient way.

Other well-performing methods are based on object shape, and they include: probabilistic graphical models (e.g., Bayesian networks [7,8]), syntactic approaches [9] or rule-based approaches [10]. Within the category of Bayesian networks methods, Kowalak [11] has proposed the characteristic points placed on the contours of objects in a scene based on NCM (Negative Curvature Minima) and PCM (Positive Curvature Maxima) points which represent the extreme of the contour. However, all these methods have the same disadvantage, namely, the number of characteristic points is too low. This results in the incorrect location of the contours of objects due to the segmentation of objects affected by the background, texture and shadow.

In this paper, for human activity recognition, we propose a novel technique based on MPEG Compact Descriptors for Visual Search (CDVS) dedicated to textural features, which are defined for entire images. CDVS is part of the MPEG-7 standard [1, 2]. This part of the MPEG-7 standard specifies an image description tool designed to enable efficient and interoperable visual search applications, allowing visual content matching in images. Visual content matching includes matching the views of objects, landmarks, and

printed documents, while being robust to partial occlusions, as well as changes in viewpoint, camera parameters, and lighting conditions [2].

The main novelty is another way of using CDVS descriptors. We propose a new type of descriptor activity as a vector of CDVS descriptors for successive images that represent human activity as a collection of poses (details are presented in Section 3). In addition, we propose the following elements: i) a new soft metric to compare both query and reference CDVS descriptors in the CDVS matching scheme after the geometric verification process, ii) a new condition for activity score to prevent incorrect classification of query descriptors which do not have corresponding subsets of activities in the reference dataset, iii) a new activity descriptor matching scheme.

There are numerous advantages of using CDVS in human activity classification. First of all, CDVS is standardized and therefore it can be assumed that many camera manufacturers will begin to equip their products with a dedicated chip to calculate CDVS in the camera. This will allow the calculation of the descriptors in real-time, in parallel to video processing and compression. The descriptors will be calculated on the source image, not after the decoding of the compressed bitstream on the server side. Some of the calculations performed so far on the side of a costly server will thus be transferred to the camera. In the case of cameras equipped with an algorithm defining CDVS descriptors, it will be possible to send only the descriptors instead of a video stream, which will result in a lower bitrate.

The goal of this paper is to show that it is possible to use standardized tools of MPEG-7 for custom applications, yielding a very high level of effectiveness in distinguishing activities, comparable to many of the most advanced methods.

The paper is organized as follows. The next section (Section 2) explains the proposed new methodology of the matching scheme, in which a new soft metric to compare both query and reference CDVS descriptors in the CDVS matching scheme is presented, followed by a novel human activity recognition scheme using CDVS descriptors, together with the definition of the proposed new activity descriptor (Section 3). Section 4 presents the results of experiments. An extensive set of experiments on two benchmark action datasets, KTH and Weizmann, have been conducted. Methods based on local textural features were chosen for comparison, e.g. 3D SIFT [5][20], SFA-based features [15], histograms of oriented 3D spatio-temporal gradients HOG3D (three dimensional Histogram of Oriented Gradients) [17], a “bag of spatial-temporal words” model combined with a space-time interest points detector [18], a "bag of words" (BOW) algorithm combined with HOG3D [19]. Section 5 provides the conclusion and the proposal for further work.

2. CDVS MATCHING SCHEME

Compact descriptors for visual search (CDVS) were designed for image retrieval applications. This standard defines the bitstream of descriptors and the descriptor extraction process. A standardized bitstream syntax of the descriptors allows for an easy integration of different databases and image retrieval services. In addition, the descriptor is scalable, which means that descriptors with different lengths can be calculated for one image. The CDVS syntax consist of two main parts which are the global descriptor and local descriptor. The global descriptor is mostly used for fast image retrieval from database. The local descriptor is used for accurate image matching and object localization. The CDVS extraction process is out of scope of this paper. More details can be found in [1].

In general, the proposed matching scheme is based on [16], with two main differences. First, the proposed solution does not use the global descriptor during the matching process. It is caused by the assumption that only images containing human actions are processed by the algorithm. Global descriptors are mostly used for a quick comparison of two different CDVS descriptors and with such an assumption, there is no need to perform this step. Second, the novel soft decision rule is introduced. This rule is able to handle the diverse nature of human actions and is described later in this section. A block diagram of the proposed matching scheme is shown in Fig. 1.

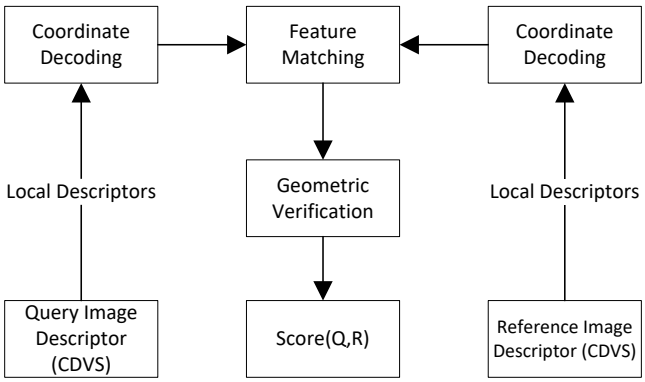

Fig. 1. Block diagram of proposed CDVS matching scheme

Let us assume that there are *Query (Q)* and *Reference (R)* CDVS descriptors. Those descriptors were calculated for two different images. The proposed solution uses only local descriptors in the matching process. Local descriptors consist of spatial coordinates of the characteristic points and the corresponding descriptor values. The spatial coordinates of characteristic points are encoded using a histogram map and histogram count and arithmetic coding. In order to match *Q* and *R* descriptors, first, the coordinates of characteristic points need to be decoded in the coordinate decoding step.

Next, the characteristic points from *Q* and *R* descriptors are matched. The matching process relies on comparing

descriptor values between *Q* and *R*. The compressed local feature descriptors are compared in the compressed domain using the L1 distance. For each matched pair of characteristic points, a match weight *w* is assigned. This weight corresponds to the importance of the match. The output of the coordinate decoding step is a collection of weighted matches P_1:

$$P_1 = ((q_1, r_1, w_1), (q_2, r_2, w_2), \dots, (q_m, r_m, w_m)) \quad (1)$$

where:

q_m, r_m – characteristic points for *m-th* match,

w_m – weight of *m-th* match.

The geometric verification step takes place after feature matching. The main goal of the geometric verification step is to detect incorrect matches between *Q* and *R*. The Distract [21] algorithm is used for verification, because it is much faster than RANSAC [12]. The output of this stage is a set of inliers matches P_2:

$$P_2 = ((q_1, r_1, w_1), (q_2, r_2, w_2), \dots, (q_i, r_i, w_i)) \quad (2)$$

where:

q_i, r_i – characteristic points for *i-th* inlier match,

w_i – weight of *i-th* inlier.

In general, the number of inliers is lower than the number of matches. Instead of making a hard decision that two CDVS descriptors match to each other, we propose a new, soft similarity metric. This soft metric eliminates the need to define the value of the threshold of compliance for two descriptors. This metric is given as the ratio of the sum of inliers weights to the sum of weights of all possible matches:

$$Score(Q, R) = \frac{\sum_{i=1}^{I} w_i}{\sum_{m=1}^{M} w_m} \quad (3)$$

where:

w_i – weight of *i-th* inlier,

w_m – weight of *m-th* match.

I, *M* – number of inliers and matches respectively.

The *Score* metric varies in the <0, 1> range. Higher values are obtained when a human pose in *Q* looks similar to a human pose in *R*. Note that the same pose looks different even if is performed by the same human twice. This metric allows us to compare different human poses in an efficient manner.

3. ACTIVITY CLASSIFICATION

Let us consider a system based on a single stationary camera that records a scene. We assume that there is only one human in the scene at once. Activities which happen in that scene on a frame level can be interpreted as a sequence of poses in time (Fig. 2). Those poses can be described using the entire picture captured by the camera. The authors propose CDVS for this purpose. As far as we know, this is the first use of CDVS for human activity classification.

Fig. 2. Activity as a sequence of poses in time.

The activity of a human can be described by a set of descriptors that represent the characteristic points, as shown in Fig. 3. This set of characteristic points at a given time instant defines a pose. In this paper, we use CDVS to detect and describe the characteristic points.

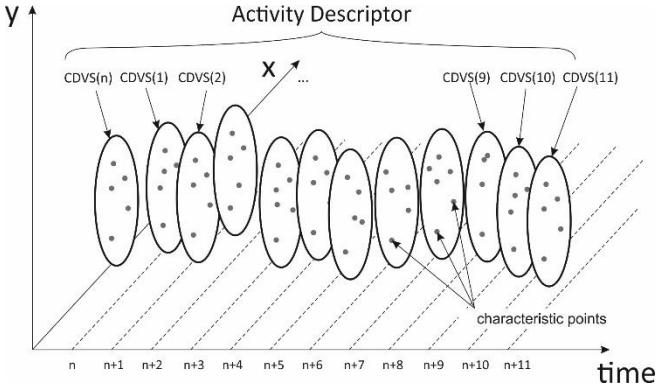

Fig. 3. Human activity as a sequence of characteristic points in time.

In general, CDVS is used to describe a single image. Human activity can be treated as a sequence of poses which correspond to consecutive image frames. In order to describe human activity, we have adapted CDVS in such a way that CDVS is calculated for each image in a video sequence. Therefore, activity descriptor *AD* is defined as a vector of CDVS descriptors and is given by:

$$AD = (CDVS(I_1), CDVS(I_2), \dots, CDVS(I_N)) \quad (4)$$

where:

I_n – *n-th* frame in sequence,

N – length of sequence.

Activity descriptors calculated for two different activities have different configurations of characteristic points. In other words, the more different is one activity from another, the more activity descriptors vary from each other. In addition, the length of activity descriptors may be different for various activities. In order to calculate the similarity score between two activity descriptors, we propose the *Score_AD* measure.

The comparison of a currently considered query activity AD_Q with reference activity AD_R is based on the accumulation of eq. 3 measure between all possible pairs of CDVS descriptors from AD_Q and AD_R sets. It is given by:

$$Score\_AD(AD_Q, AD_R) = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M} Score(AD_Q(n), AD_R(m))}{N * M} \quad (5)$$

where: $AD_Q\ (n)$ – *n-th* CDVS descriptor in the currently considered activity,

$AD_R\ (m)$ – *m-th* CDVS descriptor in the reference activity,

N,M – number of CDVS descriptors in the query and reference sets, respectively.

The *Score_AD* measure varies in <0 ,1> range. Higher value is obtained when query and reference descriptor are correctly matched. In other words, if AD_Q and AD_R correspond to two different kinds of activities, values of *Score_AD* closer to zero are expected.

It is assumed that the proposed solution is using a reference database of descriptors in the classification process. Those descriptors were calculated for different kinds of activities. The same kind of activities belong to a reference subset of activities. The classification of activity is based on the distance calculation between the currently considered descriptor and each subset of activities.

In order to classify a query activity, activity score *AS* measure is calculated for each kind of activity. First, *Score_AD* is calculated for AD_Q and each activity descriptor belonging to considered subset of activities. The results are accumulated and finally normalized by the size of considered subsets of activities. This operations are performed for each considered subset. *AS* is given by:

$$AS[i] = \frac{\sum_{l=1}^{L} Score\_AD(AD_Q, AD_R^l)}{L} \quad (6)$$

where:

i – considered *i-th* activity subset,

AD_R^l – *l-th* activity from the considered subset,

L – number of activities in the considered subset.

Finally the *AS* vector is obtained and used to create a ranking system. The size of this vector is equal to the number of considered kinds of activities in the dataset. The values in this vector correspond to the similarity of the query descriptor to each subset of activities. Those values vary in range <0,1>. Higher values are obtained if the query descriptor is similar to the reference subset of activities. The reference subset of activities which achieves the highest value of the *AS* measure is taken as a result of the classification.

The number of different kinds of activities in the reference dataset is limited. It is caused by the inability to collect test samples for all kinds of activities which may happen in a real scene. We propose a conditional check of *AS* vector values in order to prevent incorrect classification of query descriptors which do not have corresponding subsets of activities in the reference dataset. This condition is given by:

$$\begin{array}{ll} if & (\max(AS) - \min(AS)) < mean(AS) \quad query\ rejected \\ else & \quad query\ accepted \end{array} \quad (7)$$

where:

max(AS) – maximum value of *AS*,

min(AS) – minimum value of *AS*,

mean(AS) –mean value of *AS*.

In order to check that a query descriptor may belong to the reference subset of activities, the first maximum, minimum and mean values of the *AS* vector are calculated. Next, condition (7) is checked. This condition can be interpreted as follows. If the query descriptor has a corresponding subset of activities in the reference dataset, it means that there is one high value in the *AS* vector, and many low values. Otherwise, all values of this vector are low, condition (7) is fulfilled and query is rejected. Note that the whole classification process is performed without any threshold.

4. EXPERIMENTS

The main objective of the experiments was to evaluate the accuracy metric of the proposed solution. Additionally, the confusion matrix is calculated in order to examine how well the proposed solution classifies different kinds of activities. The experiments were performed on the Weizmann action dataset [14] and the KTH dataset [13]. These datasets are well known in the field of human activity recognition. Sample images from the two datasets are shown in Fig. 4 and Fig. 5, respectively.

Fig. 4. Example frames from the Weizmann dataset.

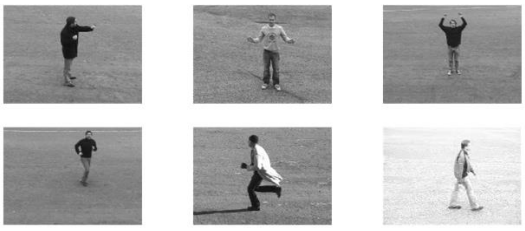

Fig. 5. Example frames from the KTH dataset.

The Weizmann dataset consists of 10 different actions: walk, run, jump, gallop sideways, bend, one-hand wave, two hands wave, jump in place, jumping jack, skip. Each action is performed by 9 actors resulting in 90 test sequences in total. Each sequence has a spatial resolution of 180 x144 points and 25 fps.

The KTH dataset contains six types of human action: boxing, handclapping, handwaving, jogging, running and

walking. Each action is performed several times by 25 actors in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. The KTH dataset contains 600 sequences with a spatial resolution of 160 x 120 points and 25 fps.

The leave one out cross validation method is used for evaluation. In this method, one actor is removed from a dataset and treated as a test sample. The rest of actors is used as training data. This procedure is repeated for each actor in the dataset. The classification decision is made based on the index of the maximum value of the *AS* vector. MPEG-7 CDVS reference software was used [16] with a 16 kB descriptor size in the experiments. The confusion matrices of classification results on the Weizmann dataset and KTH dataset are given in Table 1 and Table 2, respectively. The values in each column in this matrix correspond to the prediction rates of the considered activity type.

Table 1. Confusion Matrix on Weizmann dataset.

	Wave1	Wave2	Run	Bend	Skip	Side	Walk	Jack	Jump	PJump
Wave1	**0.89**			0.11				0.11		
Wave2		**0.89**								
Run			**0.90**							
Bend				**0.89**						
Skip					**1.00**					
Side						**1.00**				
Walk							**1.00**			
Jack	0.11	0.11	0.10					**0.89**		
Jump									**0.89**	0.11
PJump									0.11	**0.89**

Table 2. Confusion Matrix on KTH dataset.

	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	**1.00**					
Handclapping		**0.97**				
Handwaving		0.03	**1.00**			
Jogging				**0.75**	0.40	0.28
Running				0.18	**0.52**	0.05
Walking				0.07	0.08	**0.67**

The proposed approach achieves a 92.4% accuracy for the Weizmann dataset and an 81.8% accuracy for the KTH dataset. Confusions occur mostly between similar activities like running, jogging and walking, or two hand waving and jumping jack. These activities contain many similar poses which make it hard to distinguish between the types of activities. Those activities mainly differ in motion pattern which is not handled by the proposed solution. Note that in the case of treating jogging, running and walking as one kind of activity, the proposed approach achieves a 99.25% accuracy.

Table 3. Comparison of accuracy of recognition algorithms

Method	Weizmann [%]	KTH [%]
Proposed approach	**92.40**	**81.80**
Niebles[8]	90.00	83.33
Klaser[7]	84.30	91.40
Zhang[6]	93.00	-
Scovanner[5]	82.00	-
Lu[9]	93.50	91.50
Xu[10]	99.10	95.00

The comparison of the proposed method to other state-of-the-art methods is shown in Table 3. The proposed solution shows considerable accuracy even without the use of motion information on the Weizmann dataset. The Weizmann dataset consist of different kinds of activities which vary in appearance. The proposed solution performs well in such scenarios. The KTH dataset contains similar looking activities which are the main source of errors. The confusion between jogging, running and walking decreases the overall accuracy. Further improvement is possible by introducing motion information to the classification process.

5. CONCLUSION

In this paper, a novel idea for human activity recognition is presented. The proposed method uses MPEG-7 CDVS to describe a human pose and distance based ranking system for classification. In particular, we proposed the following new elements: i) a new type of descriptor activity as a vector of CDVS descriptors, ii) a new soft metric to compare both query and reference CDVS descriptors in the CDVS matching scheme after the geometric verification process, iii) a new condition for activity score to prevent incorrect classification of query descriptors which do not have corresponding subsets of activities in the reference dataset, iv) a new activity descriptor matching scheme. The whole classification process takes place without the use of any threshold.

The experiments were performed on the Weizmann and KTH datasets. The proposed method achieves a 92.4% accuracy for Weizmann dataset and an 81.8% accuracy for KTH dataset. The proposed method shows a weakness in distinguishing between jogging, running and walking. These activities contain many similar poses which make it hard to distinguish between the types of activities. The presented method can correctly distinguish between activities which consist of different human poses. Similar-looking activities are still a problem, because they differ only in motion, which is not described by CDVS.

The experimental results reported in this paper demonstrate limited efficiency of the proposed method. The reason of such efficiency is related to the lack of motion information exploited in the proposed method which is

currently in the first stage of research. Currently, we are working towards exploitation of motion information in the human activity recognition that will yield higher efficiency of the techniques that employs standard CDVS compact descriptors. Such descriptors will allow interoperability of hardware and software provided by many manufacturers. The intelligent surveillance cameras would be possible to be quickly installed everywhere where needed, e.g. in the house of a person coming back home after surgery etc. As the calculation of the CDVS descriptor would be done in the camera, cheap wireless transmission of the compact descriptor would be the main communication need. Moreover, the employment of standard descriptors will likely result in cost decrease, as the same video analysis will be done in many surveillance applications, both in security sector and, likely, in e-health sector. Therefore, the authors believe in the importance of the research presented.

ACKNOWLEDGMENT

Research project was supported by public funds under Project 08/84/DSPB/0190 Poznań University of Technology, Chair of Multimedia Telecommunications and Microelectronics.

REFERENCES

[1] L.-Y. Duan et al., “Overview of the MPEG-CDVS standard,” IEEE Trans. Image Process. , vol. 25, no. 1, pp. 179–194, Jan. 2016.

[2] ISO/IEC IS 15938-13, Information Technology—Multimedia Content Description Interface—Part 13: Compact Descriptors for Visual Search.

[3] Laptev I., Marszalek M., Schmid C., and Rozenfeld B., “Learning realistic human actions from movies,” in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, pp. 1–8., 2008.

[4] Mikolajczyk K. and Schmid C., “Scale & Affine Invariant Interest Point Detectors,” International Journal of Computer Vision, vol. 60, no. 1, pp. 63–86, 2004.

[5] P. Scovanner, S. Ali, and M. Shah, "A 3-Dimensional SIFT Descriptor and its Application to Action Recognition," in Proceedings of the 15th international conference on Multimedia, pp. 357-360, 2007.

[6] Bay H., Tuytelaars T., and Van Gool L., “Surf: Speeded up robust features,” in Computer Vision–ECCV 2006, pp. 404–417. Springer, 2006.

[7] H. Buxton and S. Gong, "Visual surveillance in a dynamic and uncertain world", Artificial Intelligence, 78(1-2):431–459, 1995.

[8] S. S. Intille and A. F. Bobick. "A framework for recognizing multi-agent action from visual evidence", AAAI Conference on Artificial intelligence, pages 518–525, 1999.

[9] Y. A. Ivanov and A. F. Bobick, 'Recognition of visual activities and interactions by stochastic parsing", IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8):852–872, 2000.

[10] H. Dee and D. Hogg, "Detecting inexplicable behaviour", British Machine Vision Conference, pages 477–486, 2004.

[11] Kowalak K., Kaminski L., Gardzinski P., Mackowiak S., Hofman R., "Human Behavior Recognition Using Negative Curvature Minima and Positive Curvature Maxima Points", Computer Communication Networks and Telecommunications, New Research in Multimedia and Internet Systems: Vol. 314, pp. 57-66, 2015.

[12] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6):381– 395, 1981.

[13] C. Schüldt, I. Laptev, B. Caputo "Recognizing human actions: a local svm approach", Pattern Recognition, ICPR 2004. Proceedings of the 17th International Conference on. IEEE, vol. 3, pp. 32-36, 2004.

[14] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri "Actions as pace-time shapes", Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 12, pp. 2247-2253, 2007.

[15] Z. Zhang and T. Dacheng, "Slow Feature Analysis for Human Action Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, pp. 436-450, 2012.

[16] ISO/IEC DIS 15938-14 Reference software, conformance and usage guidelines for compact descriptors for visual search.

[17] A. Klaser and M. Marszalek, "A spatio-temporal descriptor based on 3D-gradients," presented at the BMVC, 2008.

[18] C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," International journal of computer vision, vol. 79, pp. 299-318, 2008.

[19] M. Lu, L. Zhang, “Action recognition by fusing spatial-temporal appearance and the local distribution of interest points,” in 2014 International Conference on Future Computer and Communication Engineering (ICFCCE 2014). Atlantis Press, 2014.

[20] K. X. Xinghao, J. T. Sun “Human activity recognition based on pose points selection” IEEE International Conference on Image Processing (ICIP), pp.2930-2834 2015.

[21] S. Lepsøy, G. Francini, G. Cordara, and P. P. B. de Gusmao, “Statistical modelling of outliers for fast visual search,” in Proc. IEEE Workshop Vis. Content Identificat. Search (VCIDS), Barcelona, Spain, Jul. 2011,pp. 1–6.

[22] Lertniphonphan K., Aramvith S., Chalidabhongse T.H., “Human Action Recognition using Direction Histograms of Optical Flow”, 11th International Symposium on Communications and Information Technologies