HUMAN ACTION RECOGNITION WITH MPEG-7 DESCRIPTORS AND ARCHITECTURES

Zia Moghaddam
University of Technology, Sydney
Broadway, Ultimo NSW 2007, Australia
ziam@it.uts.edu.au

Massimo Piccardi
University of Technology, Sydney
Broadway, Ultimo NSW 2007, Australia
massimo@it.uts.edu.au

ABSTRACT

Modern video surveillance requires addressing high-level concepts such as humans' actions and activities. In addition, surveillance applications need to be portable over a variety of platforms, from servers to mobile devices. In this paper, we explore the potential of the MPEG-7 standard to provide interfaces, descriptors, and architectures for human action recognition from surveillance cameras. Two novel MPEG-7 descriptors, symbolic and feature-based, are presented alongside two different architectures, server-intensive and client-intensive. The descriptors and architectures are evaluated in the paper by way of a scenario analysis.

Categories and Subject Descriptors

I.2.10 [**Vision and Scene Understanding**]: *Architecture and control structures.*

General Terms

Standardization

Keywords

Human action recognition, MPEG-7, MPEG-7 visual descriptors, Client-server architecture.

1. INTRODUCTION

Automated video surveillance aims to detect objects and events of interest for safety and security via automated analysis of closed-circuit television (CCTV) videos. As an area of technology, automated video surveillance is still relatively recent and dominated by many proprietary solutions that are neither standard nor modular. Consequently, integration of products from different manufacturers is still in itself an arduous and onerous task. We envisage that video surveillance experience the same adoption of standardised interfaces that has benefited other areas of technology such as telecommunications and computer networks in the past.

Certain multimedia standards such as MPEG-7 (the standard for multimedia content description from the ISO MPEG committee) could be used for the standardisation of interfaces in video surveillance applications [15]. MPEG-7 was originally proposed by the MPEG committee for facilitating efficient search and retrieval from multimedia databases. Being an open standard, it could be used in a variety of video surveillance applications including moving object detection and classification, object tracking, human action recognition and others.

Since the emergence of the MPEG-7 standard, various papers have been published exploiting MPEG-7 visual descriptors for video surveillance applications. One of the first works, from Berriss *et al.*, dates 2003 [2]. In this work, the authors exploited two MPEG-7 descriptors for tracking people entering and exiting a store monitored by a camera. Goldmann *et al.* in [9] proposed recognising human postures based on an MPEG-7 shape descriptor and a feature vector derived from projection histograms. Annesly *et al.* in [1] evaluated retrieval efficiency of four MPEG-7 colour descriptors - Dominant Colour, Colour Layout, Colour Structure and Scalable Colour - alongside two own introduced validation descriptors - Mean (r,g,b) and random - for matching people entering and exiting a room monitored by two distinct cameras. Chien *et al.* in [6] proposed a new MPEG-7 descriptor, named HCSD (Human Colour Structure Descriptor), for detecting humans in surveillance videos.

In recent times, human action recognition (HAR) has become a very intensive research area of video surveillance for its potential use in a variety of security and safety applications. HAR addresses the automated classification of a sequence of frames depicting a human action into one of several, pre-defined classes. According to the reviewed literature, previous research has mainly exploited MPEG-7 for classification of still frames rather than that of whole frame sequences. However, human action recognition adds the new dimension of time and requires the extension of existing visual descriptors to sequences of frames. We argue in the next section that the various current MPEG-7 shape, colour and motion descriptors do not satisfy the requirements of human action recognition. Hence, as the first contribution of this paper, we propose two new MPEG-7 descriptors specific for human action recognition. Furthermore, two novel, practicable MPEG-7 based architectures are introduced for human action recognition under real-time constraints and over a variety of platforms such as servers, PCs and mobile devices.

This paper is organized as follows: section 2 offers a brief introduction to the elements of the MPEG-7 standard required for the following discussion, and discusses the inadequacies of current MPEG-7 descriptors for the task of human action recognition. Section 3 articulates the main steps of human action recognition. In section 4, we propose and compare two MPEG-7 based architectures alongside two new descriptors for human action recognition. Finally, conclusions are presented.

2. BASICS OF THE MPEG-7 STANDARD

In 2001, the Moving Picture Experts Group (MPEG) proposed the MPEG-7 standard for describing multimedia contents as metadata and enabling effective searching, filtering and indexing in a multimedia database. The standard contains a set of *descriptors* representing different multimedia features, including visual and audio descriptors [15]. The MPEG-7 visual descriptors include colour, texture, shape, motion, localization and face recognition descriptors [15]. The power of MPEG-7 is its extendibility using the Data Definition Language (DDL) which allows the creation of new descriptors [15].

The MPEG-7 Reference Software, also called eXperimentation Model (XM), is used to validate MPEG-7 descriptors. XM provides applications for extracting descriptors from the input media database and storing them in description files. In addition, client applications in XM use the extracted MPEG-7 descriptors for retrieval, filtering or transcoding tasks based on a query [15]. For example, the search-and-retrieval client application in Figure 1 first loads the descriptors from a descriptor database file. The input query is then matched against all the descriptors to produce a sorted list of matching media with decreasing similarity to the query.

To extend MPEG-7 descriptors to video classification - such as in human action recognition - two main approaches could be considered. In the first approach, the video would be processed as a sequence of frames and, consequently, in each frame shape, colour and/or texture descriptors of each human subject would be exploited. In the second approach, we should consider the video as a whole unit and motion descriptors would be the appropriate choice. Hereafter, we briefly review such descriptors.

2.1 MPEG-7 descriptors for human action recognition

Shape information is one of the most useful properties to inform human action recognition. Whereas a variety of specific shape features have been used in the sector literature (contours, snakes, interest points, region templates and others), contour-based shape features have proved discriminative and computationally lightweight [5, 8]. MPEG-7 contains a contour-based shape descriptor [4, 19] based on the curvature scale space (CCS) of an object's outline. However, CCS was intentionally designed to be invariant to rotation and therefore is not suitable to discriminate across certain types of human actions.

Among its visual motion descriptors, MPEG-7 enlists a Motion Activity Descriptor [10, 19]. This descriptor denotes the overall intensity of an action (for instance, slow or fast paced) and is too simplistic to fully describe human actions. The Motion Trajectory Descriptor is another MPEG-7 motion descriptor which characterises the displacement of a representative point of each moving object over the time [10, 19]. The current implementation of this descriptor exploits the spatio-temporal positions of the object's centroid [10] and is not suitable to encode the articulated shape of humans. Moreover, as we discuss in the next section, the use of MPEG-7 Motion Trajectory Descriptor may imply a considerable overhead due to the uncertain time segmentation of human actions.

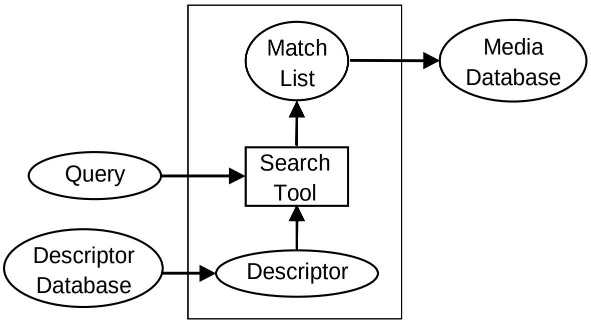

Figure 1. Use of descriptors for media retrieval in a client application.

3. HUMAN ACTION RECOGNITION

Human action recognition is a high-level video analysis task relying on several, lower-level tasks such as object segmentation, tracking and posture recovery. The typical goal of automatic action recognition is the classification of a given frame sequence depicting a single object as one of several classes of pre-defined actions.

The main challenge in HAR is the significant diversity among various instances of the same action performed by different people. Moreover, every individual performs each action in a different manner over various instances, both in space and time [16]. As a consequence, the within-class variance tends to be large and class separation correspondingly small, challenging accurate classification. In addition, HAR is hindered by feature extraction inaccuracies due to occlusions, changes in illuminations and the deformable nature of human subjects.

An MPEG-7 based human action recognition system should respond to an MPEG-7 query for retrieval of a specific action. The related query could be in the following general format: "*Show me the frame sequences of people performing action <ActionClass> during interval [T_{start} … T_{end}] in the areas inspected by cameras {C_1, …, C_M}*". In the following, we explain how action recognition is sub-divided into main steps and how they relate to the query.

3.1 Human action recognition steps

In general, any action recognition approach consists of two main steps: 1) the extraction of a feature set from the video data and 2) action classification based on the extracted features. However, other processing steps are possibly involved such as: 3) foreground extraction, 4) tracking and 5) time segmentation.

Step 1- Feature extraction: A variety of features were suggested for human action recognition such as optical flow [7], body parts' tracking [20], silhouette-based approaches [3, 8, 13], spatio-temporal feature descriptors (cuboids, HOG/HOF, HOG3D, extended SURF and others) [22]. However, researchers are left with the decision whether to use a rich feature set, possibly invariant to the viewpoint (e.g. [18]), or a simple, fast-to-extract feature set designed with opportunistic action discrimination in mind [5, 8, 17].

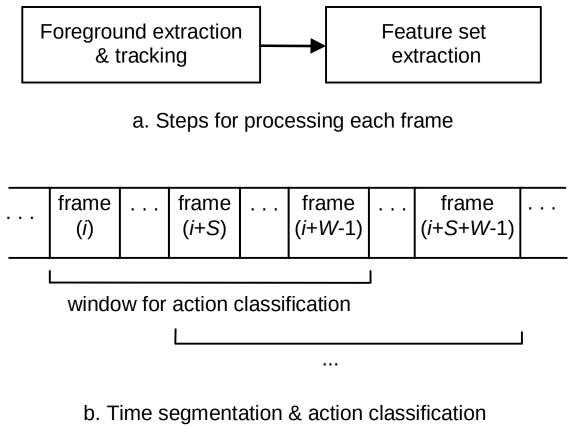
Foreground extraction & tracking
Feature set extraction
a. Steps for processing each frame
. . .
frame (*i*)
frame (*i*+*S*)
frame (*i*+*W*-1)
frame (*i*+*S*+*W*-1)
window for action classification
...
b. Time segmentation & action classification

Figure 2. Human action recognition steps.

Step 2- Action classification: The main approaches to action classification are based on template matching in the domain of time such as dynamic time warping (DTW) or graphical models such as the hidden Markov model (HMM) and other dynamic Bayesian networks. Discriminative models such as conditional random fields have also been used extensively [23]. In recent years, also simple classification of histogram features collected over space-time grids has been applied successfully [11].

Step 3- Foreground extraction: In some cases, the extraction of the feature set is preceded by the extraction of the image's *foreground pixels*. While this step is not strictly necessary for action recognition and is regarded as a potential source of early errors, it is still often applied in applications where reliable background modelling is possible. The extraction of the foreground pixels can help identify regions of meaningful features and solve the data association problem in the presence of multiple actors. Hence, this preliminary step returns the foreground masks of candidate objects (blobs) in each frame.

Step 4- Tracking: Tracking and data association need to be performed to track each single object along the frame sequence. In some approaches, the entire object is tracked at once, in others individual features are tracked explicitly and the object's location is inferred [24].

Step 5– Time segmentation: Prior to attempting classifications, the start and end times, T_{start}, T_{end}, of an action should be determined (time segmentation). Depending on the specific scenario, information may be available to support time segmentation prior to recognition. In some cases, action classification and time segmentation have been attempted jointly [21]. Very often, however, time segmentation is conducted in terms of fixed-length, overlapped windows of frames. The length of the window, *W*, and the stride between windows, *S*, must permit an approximate alignment with the action's actual time segment [5].

Figure 2 illustrates the various processing steps of action recognition. Figure 2b also gives evidence to the main drawback of descriptors such as the MPEG-7 Motion Trajectory Descriptor or similar when used for human action recognition. The main problem lies in the high data redundancy deriving from the overlapped windows: with these descriptors, the same frame description would be repeated *W*/*S* times. With practical values for this ratio (in the order of 5÷10), the redundancy is excessive.

4. MPEG-7 BASED HUMAN ACTION RECOGNITION ARCHITECTURES

In this section, we introduce two MPEG-7 based architectures, namely, *Server-Intensive* and *Client-Intensive*, to perform the human action recognition steps discussed in section 3.1. These architectures mainly follow the model of MPEG-7 extraction and search applications, discussed in section 2.

In both architectures, we allocate the computationally-intensive tasks of foreground extraction, tracking and feature set extraction to a so-called *server system*. Conversely, the client device is a system responsible for the search and retrieval tasks. Moreover, depending on the processing power and connection bandwidth of the client device, we categorize it as either a *thick client*, such as a PC, or a *thin client*, such as a Personal Digital Assistant (PDA), a mobile device or a smartphone.

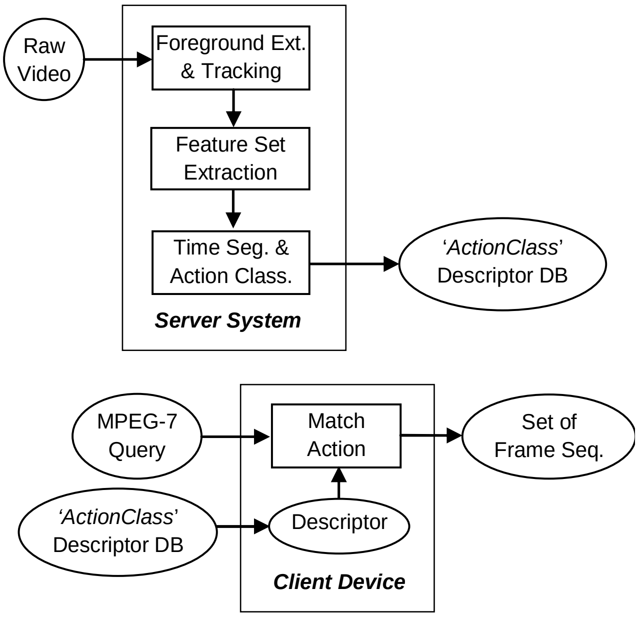

Figure 3. ***Server-Intensive*** **architecture.**

4.1 *Server-Intensive* architecture

In the case of a thin client device, we suggest the *Server-Intensive* architecture would offer a good resources' balance (Figure 3). In this architecture, the server system performs all the steps of foreground extraction, tracking, feature set extraction, time segmentation and action classification. It then stores the recognized action for a sequence of frames in the form of a text string in a new motion descriptor which we called the *ActionClass* descriptor. The client device performs just the lightweight task of MPEG-7 query matching against all the *ActionClass* descriptors in the descriptor database. An example of the *ActionClass* motion descriptor is shown in Figure 4.

```
<Descriptor xsi:type="ActionClass">
     <FrameSeqStart> 400 </FrameSeqStart>
     <FrameSeqEnd> 450 </FrameSeqEnd>
     <BlobIdentifier> 13 </BlobIdentifier>
     <ActionName> “Run” </ActionName>
</Descriptor>
```

Figure 4. An example of the *ActionClass* motion descriptor.

4.2 *Client-Intensive* architecture

In contrast with the *Server-Intensive* architecture, we suggest the *Client-Intensive* architecture would offer an appropriate resource tradeoff for thick clients (Figure 5). Here, the server system only performs the foreground extraction, tracking and feature set extraction steps for each frame and each object, and then stores the extracted features in a new visual descriptor which we called *ObjectFeatures*. On the client side, the client device uses the *ObjectFeatures* descriptors for time segmentation and action classification, and performs the MPEG-7 query matching. With this architecture, the client device has the further freedom of choosing a time segmentation and action classification approach of its own choice.

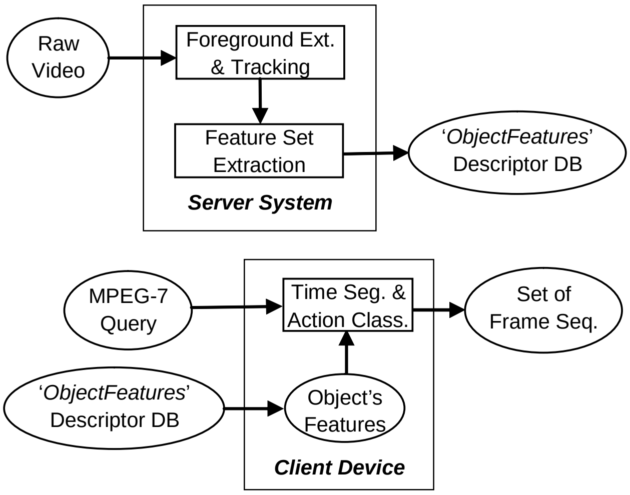
Raw Video
Foreground Ext. & Tracking
Feature Set Extraction
Server System
‘ObjectFeatures’ Descriptor DB
MPEG-7 Query
Time Seg. & Action Class.
Set of Frame Seq.
‘ObjectFeatures’ Descriptor DB
Object’s Features
Client Device

Figure 5. *Client-Intensive* architecture.

The *ObjectFeatures* visual descriptor should ideally be simple and fast to extract to satisfy the requirements of real-time human action recognition. Here, we illustrate two examples for this descriptor. The first is based on the sectorial extreme points of [17] which describe the position of physical points such as head, left and right hands and feet in the object’s silhouette. Based on [17], the actor’s silhouette is divided into five circular sectors centred around its centroid. Then, for each sector the silhouette’s contour point farthest from the centroid is determined. Further, to also encode the absolute position of the object, the centroid’s coordinates are added to the feature set. Figure 6 depicts the extracted points for one frame of the ‘Jumping-jack’ action from the Weizmann video dataset [3]. The corresponding *ObjectFeatures* descriptor is shown in Figure 7.

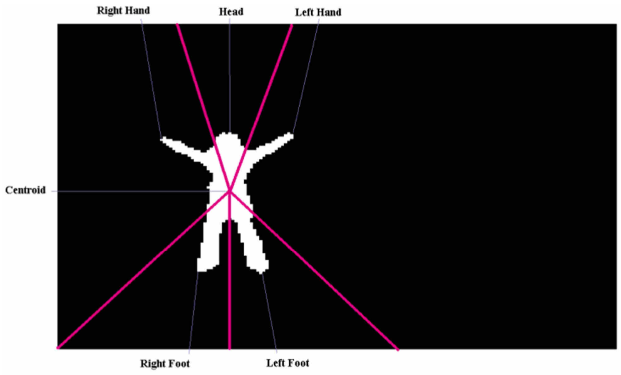

Figure 6. Extracted sectorial extreme points for one frame of the Weizmann video dataset.

```
<Descriptor xsi:type="ObjectFeatures_ExtPts">
     <FrameNo> 11 </FrameNo>
     <BlobIdentifier> 1 </BlobIdentifier>
     <CentroidCoords > 77 56 </CentroidCoords>
     <HeadCoords> 49 54 </HeadCoords>
     <LHandCoords> 49 75 </LHandCoords>
     <RHandCoords> 51 35 </RHandCoords>
     <LFootCoords> 111 66 </LFootCoords>
     <RFootCoords> 110 46 </RFootCoords>
</Descriptor>
```

Figure 7. An example of the *ObjectFeatures* visual descriptor using the sectorial extreme points of [17].

As another example of *ObjectFeatures* visual descriptor, we exploit the projection histograms as the feature set. Projection histograms consist of the frequency bins of an object’s pixels projected onto the image coordinate axes [9]. As an action takes place, the two projection histograms reflect the changes in the object’s shape and can be used for action discrimination. We use histograms with 10 bins each, leading to a total feature set of size F = 20. Figure 8 depicts the projection histograms of the same frame of figure 6. The resulting *ObjectFeatures* descriptor is shown in figure 9.

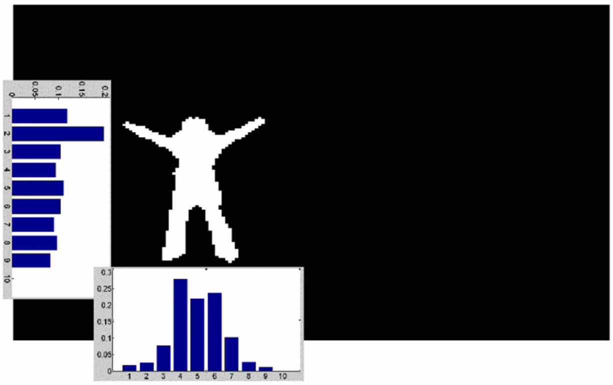

Figure 8. Projection histogram features for one frame of the Weizmann video dataset.

```
<Descriptor xsi:type="ObjectFeatures_ProjHist">
   <FrameNo> 11 </FrameNo>
   <BlobIdentifier> 1 </BlobIdentifier>
   <HorizontalBins>
      0.119  0.197  0.104  0.094  0.111
      0.104  0.090  0.097  0.082  0.000
   </HorizontalBins>
   <VerticalBins>
      0.017  0.026  0.077  0.279  0.221
      0.238  0.102  0.027  0.011  0.000
   </VerticalBins>
</Descriptor>
```

Figure 9. An example of the *ObjectFeatures* visual descriptor using projection histograms as the feature set.

The two examples of the *ObjectFeatures* visual descriptor show that this descriptor is not restricted to specific feature sets. The two fundamental identifiers are those of the frame and the object: the identification of the frame is implicit and that of the object is unavoidable if the feature set is to be referred to a specific subject. Moreover, the frame identifiers do not need to be contiguous and permits sparse frame encoding as with spatio-temporal interest points [22].

4.3 Architecture evaluation

In this subsection, we comparatively evaluate the two proposed architectures with thick and thin clients. The evaluation mainly depends on the execution time of the time segmentation and action classification steps since the other steps are expected to be executed on a server of theoretically unrestricted resources. The parameters involved in the evaluation are:

- $P = (T_{end} - T_{start})$: the observation period of interest, in seconds;
- O: the average number of objects in each frame;
- W: the window size, in frames;
- S: the stride size, in frames;
- F: the frame rate, in frames per second;
- X: the average running time of action classification, in seconds per frame per object

Each frame is processed (W/S) times due to overlapping windows, hence, the total execution time for action classification (R) is given by:

$$R = P \cdot F \cdot X \cdot O \cdot W / S \qquad (1)$$

To evaluate the average computational time for action classification we have performed a human action recognition experiment on the Weizmann video dataset [3]. In the experiment, we have classified 93 videos for a total of 11,374 frames. The execution time with a Matlab implementation on an Intel Core 2 CPU at 2.0 GHz was 6.9 s which corresponds to 0.606 ms per frame. Assuming the implementation could be optimised in a more efficient language such as C for, say, a 10-time speedup, it would be possible to reduce X to approximately 0.0606 ms. Hence, for a possible scenario with a sequence of duration P = 10 minutes, O = 8 moving objects in the scene on average, window W = 50 frames, stride S = 10 frames and frame rate F = 25 fps, the execution time R would become 36.4 seconds. This response time is typical of a thick client in a *Client-Intensive* architecture and can be regarded as acceptable and “real time”. However, with the same architecture but a thin client such as a mobile handset 10 times slower than a typical PC, the response time would become over 6 minutes and obviously not acceptable in real-time surveillance applications. Consequently, in the case of a thin client, the alternative *Server-Intensive* architecture should be considered as the solution of choice.

5. CONCLUSIONS

In this paper, we have discussed the importance of using standardised interfaces in video surveillance applications and in particular in human action recognition. We argue that the existing MPEG-7 visual descriptors are not adequate for the task and that new descriptors are needed. In the paper, two novel MPEG-7 based architectures, namely Server-Intensive and Client-Intensive, have been proposed alongside two new descriptors, the *ActionClass* motion descriptor and the *ObjectFeatures* visual descriptor. We conclude that the Server-Intensive architecture is the most appropriate in the case of “thin” client devices such as PDAs and mobile phones, whereas the Client-Intensive architecture is the most suitable when “thick” clients such as desktops can be employed. The performance analysis presented in this paper can also be parametrised to specific platforms and the approach outlined can be extended to other architectures such as server-less architectures and intelligent cameras [12, 14].

6. REFERENCES

[1] Annesley, J., Orwell, J. and Renno, J.-P. 2005. Evaluation of MPEG7 Color Descriptors for Visual Surveillance Retrieval. In *Proceedings of the 2nd Joint IEEE Intl. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance,* (Beijing, China, Oct. 15-16, 2005), 105-112.

[2] Berriss, W. P., Price, W. G. and Bober, M. Z. 2003. The Use of MPEG-7 for Intelligent Analysis and Retrieval in Video Surveillance. In *Proceedings of the IEE Symposium on Intelligence Distributed Surveillance Systems,* (London, UK, Feb. 26, 2003), 8/1-8/5.

[3] Blank, M., Gorelick, L., Shechtman, E., Irani, M. and Basri, R. 2005. Actions as Space-Time Shapes. In *Proceedings of the 10th IEEE Intl. Conf. on Computer Vision (ICCV) Volume 2,* (Beijing, China, Oct. 17-20, 2005), 1395-1402.

[4] Bober, M. 2001. MPEG-7 visual shape descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11, 6 (Jun. 2001), 716-719.

[5] Chen, H.-S., Chen, H.-T., Chen, Y.-W. and Lee, S.-Y. 2006. Human action recognition using star skeleton. In *Proceedings of the 4th ACM Intl. Workshop on Video Surveillance and Sensor Networks,* (Santa Barbara, California, USA, Oct. 23-27, 2006), 171-178.

[6] Chien, S.-Y., Chan, W.-K., Cherng, D.-C. and Chang, J.-Y. 2006. Human Object Tracking Algorithm with Human Color Structure Descriptor for Video Surveillance Systems. In *Proceedings of the IEEE Intl. Conf. on Multimedia and Expo (ICME),* (Toronto, Canada, Jul. 9-12, 2006), 2097-2100.

[7] Efros, A. A., Berg, A. C., Mori, G. and Malik, J. 2003. Recognizing action at a distance. In *Proceedings of the 9th IEEE Intl. Conf. on Computer Vision (ICCV) Volume 2,* (Nice, France, Oct. 13-16, 2003), 726-733.

[8] Fujiyoshi, H. and Lipton, A. J. 1998. Real-time human motion analysis by image skeletonization. In *Proceedings of the 4th IEEE Workshop on Applications of Computer Vision,* (Princeton, New Jersey, USA, Oct. 19-21, 1998), 15-21.

[9] Goldmann, L., Karaman, M. and Sikora, T. 2004. Human Body Posture Recognition Using MPEG-7 Descriptors. *Visual Communications and Image Processing,* vol. 5308 of *Proceedings of SPIE,* (San Jose, California, USA, Jan. 2004), 177-188.

[10] Jeannin, S. and Divakaran, A. 2001. MPEG-7 visual motion descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11, 6 (Jun. 2001), 720-724.

[11] Laptev, I., Marszałek, M., Schmid, C. and Rozenfeld, B. 2008. Learning realistic human actions from movies. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR),* (Anchorage, Alaska, Jun. 24-26, 2008), 1-8.

[12] Lee, J. Y. B. and Leung, R. W. T. 2002. Study of a serverless architecture for video-on-demand applications. In *Proceedings of the IEEE Intl. Conf. on Multimedia and Expo (ICME) Volume 1,* (Lausanne, Switzerland, Aug. 26-29, 2002), 233-236.

[13] Li, N. and Xu, D. 2008. Action recognition using weighted three-state Hidden Markov Model. In *Proceedings of the 9th Intl. Conf. on Signal Processing (ICSP),* (Beijing, China, Oct. 26-29, 2008), 1428-1431.

[14] Li, W., Kharitonenko, I., Lichman, S. and Weerasinghe, C. 2006. A Prototype of Autonomous Intelligent Surveillance Cameras. In *Proceedings of the IEEE Intl. Conf. on Advanced Video and Signal Based Surveillance (AVSS),* (Sydney, Australia, Nov. 22-24, 2006), 101-101.

[15] Martínez, J. M., 2004, *MPEG-7 Overview*, Last updated Oct. 2004, http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm.

[16] Moeslund, T. B., Hilton, A. and Krüger, V. 2006. A Survey of Advances in Vision-Based Human Motion Capture and Analysis. *Computer Vision and Image Understanding*, 104, 2-3 (Nov. - Dec. 2006), 90-126.

[17] Moghaddam, Z. and Piccardi, M. 2009. Deterministic Initialization of Hidden Markov Models for Human Action Recognition. In *Proceedings of the Digital Image Computing: Techniques and Applications (DICTA),* (Melbourne, Australia, Dec. 1-3, 2009), 188-195.

[18] Shen, Y. and Foroosh, H. 2008. View-invariant action recognition using fundamental ratios. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR),* (Anchorage, Alaska, Jun. 24-26, 2008), 1-6.

[19] Sikora, T. 2001. The MPEG-7 Visual Standard for Content Description—An Overview. *IEEE Transactions on Circuits and Systems for Video Technology*, 11, 6 (Jun. 2001), 696-702.

[20] Song, Y., Goncalves, L. and Perona, P. 2003. Unsupervised learning of human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 25, 7 (Jul. 2003), 814-827.

[21] Vezzani, R., Piccardi, M. and Cucchiara, R. 2009. An efficient Bayesian framework for on-line action recognition. In *Proceedings of the 16th IEEE Intl. Conf. on Image Processing (ICIP),* (Cairo, Egypt, Nov. 7-10, 2009), 3553-3556.

[22] Wang, H., Ullah, M. M., Klaser, A., Laptev, I. and Schmid, C. 2009. Evaluation of local spatio-temporal features for action recognition. In *Proceedings of the British Machine Vision Conf. (BMVC),* (London, UK, Sep. 7-10, 2009), 127-138.

[23] Wang, L. and Suter, D. 2007. Recognizing Human Activities from Silhouettes: Motion Subspace and Factorial Discriminative Graphical Model. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR),* (Minneapolis, Minnesota, USA, Jun. 18-23, 2007), 1-8.

[24] Yilmaz, A., Javed, O. and Shah, M. 2006. Object tracking: A survey. *ACM Compuing Surveys*, 38, 4 (Dec. 2006), 1-45.