

Study on Human Behaviour Retrieval

Yan CHEN

Faculty of Information
Technology
University of Technology,
Sydney
Sydney, NSW, Australian

Qiang WU

Faculty of Information
Technology
University of Technology,
Sydney
Sydney, NSW, Australian

Xiangjian HE

Faculty of Information
Technology
University of Technology,
Sydney
Sydney, NSW, Australian

Abstract Human behavior analysis is a hot topic in computer vision and is applied widely in many applications. Human behavior retrieval is another frontier technology in the area of multimedia information retrieval, which is related to human behavior analysis but holds several differences because of its special application purpose. Human behaviour retrieval to some extent is similar to human behaviour analysis, but the technology used for human behavior analysis cannot be used for human behavior directly. This paper will address such kind of differences and review several technologies including video retrieval, feature extraction, similarity measure and human behavior analysis. This paper will also address the importance of human behaviour retrieval. The ideas unveiled by this paper will benefit the research community and indicate a direction of human behavior retrieval research

Keywords: Behaviour retrieval, Behaviour analysis

1 Introduction

Human behavior analysis is receiving increasing attention in the area of computer vision and it has been applied in many areas, such as athletic performance analysis, surveillance and so on. There have been many such systems. For example, the real time W⁴ [1] detects and tracks groups of people as well as monitors their behaviors even in the presence of occlusion and in outdoor environments. In [2], Zhu et. al proposed a system to recognize the action of tennis player and hence to improve player's performance. The metric currently used for evaluation of tools for human behavior analysis is recognition rate. The higher recognition rate is, the better the corresponding tool is regarded to be. However, such kind of metric is not enough for human behavior retrieval purpose which needs another significant metric called recall rate.

With the development of digital libraries, retrieval of an image or a clip of video from a video database becomes more and more difficult. Traditionally, keywords are used as text labels for quickly accessing large quantity of visual data. Some search engine giants such as Google and Yahoo

also use meta-data information as keywords for image/video retrieval. The representation of visual data using text labels requires a large amount of manual work, which is not efficient and time-consuming. Therefore, it is important to automatically interpret image/video in order to save tremendous human efforts. Content-Based Image/Video Retrieval (CBIR/CBVR) is a solution to the above problems. CBIR/CBVR which retrieves targets based on visual information such as color, texture and shape has been studied in the previous years [3, 4]. It has shown many applications in medical science, art gallery and military. Human behavior retrieval [3] is a new research topic in the image retrieval area. The existing approaches to human behavior retrieval require domain knowledge and other objects information. For example, for behavior retrieval in a tennis game, the information about tennis balls is often used. Furthermore, current methods for human behavior retrieval can only be applied to specific areas [3]. Although human behavior retrieval is similar to human behavior analysis in some aspects, the methods used on human behavior analysis cannot be directly applied for human behavior retrieval.

The remaining sections are organized as follows. Existing work on general video retrieval development, content based feature extraction, similarity measurement, and preliminary research on human behavior retrieval are reviewed in Section 2 to Section 5. Section 6 indicates the possible future work related to human behavior retrieval. It is concluded in Section 7.

2 Recent Research Development on Video Retrieval and Annotation

Video is rapidly becoming the most popular media due to its high information and entertainment power. Applications that benefit from video are for education and training, marketing support, entertainment, sports etc. A straightforward approach to video retrieval is to represent visual contents in textual form (e.g. keywords). These keywords serve as indices to access the associated visual data and can be obtained when subtitle or transcript exist. The keyword approach has the advantage that visual database can be accessed using standard query language such as SQL language. However, sometimes this needs lots of extra manual processing. Therefore, there has been a

new focus on developing content-based video retrieval/annotation system. Content based video retrieval is regarded as an extension of content-based image retrieval. Moreover, compared with image retrieval video retrieval features in several factors. These factors are primarily related to the temporal information available from a video document. While these factors may complicate the querying system, they may help in characterizing useful information for the querying. The temporal information firstly induces the concept of motion of the objects presented in the document.

Generally, content based video retrieval includes three parts that are segmentation, indexing, and query processing. Segmentation divides the video into shots or scenes, and selects one or more key frames for each shot. A shot is a set of contiguous frames all acquired through a continuous camera recording. The partitioning of the video into shots generally does not refer to any semantic analysis. Only the temporal information is used. Video shot cut involves identifying the frames where a transition takes place from one shot to another.

In the next step, features are extracted from key frames supplied by the segmentation process and used to create a database index. When a query comes, segmentation and key frame extraction are performed. Then, the necessary features are extracted from the key frames according to the query. A general video retrieval flow is shown in Fig 1

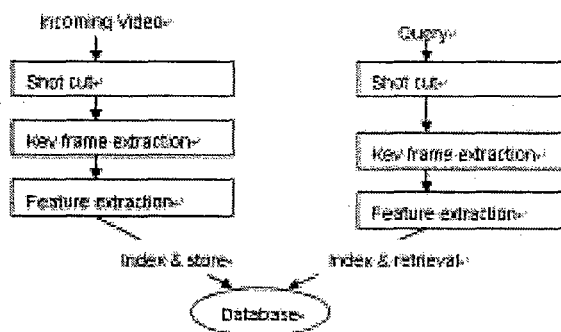


Fig. 1. Flow of content based video retrieval.

Most of the general video retrievals are based on low level features such as color, texture, shape, and motion information. Although they can be applied to a wide variety of generic video, those visual features have a common drawback that they can represent only low-level information. Some specific video retrieval approaches such as those used for improving sports performance [3] combined domain knowledge with low level features. This type of retrieval which uses the technology of human behavior analysis will have high accuracy, but its drawback is that it can be used only in a specific area and to achieve large amount of domain knowledge. But the aim of the video retrieval is not only high accuracy, but also fast

speed and high efficiency through dramatically reducing the amount of unrelated video data.

3 Contents Based Feature Extraction

Image representation is the first step for image and video retrieval. Most of image databases have been preprocessed to obtain image features such as color, texture and shape for retrieval. What feature to be used for image/video retrieval depends on applications. Because feature extraction is an important step in image and video retrieval, we review the work on extraction of color, texture and shape as follows.

3.1 Color

Color is perhaps the most dominant and distinguishing visual feature and is one of the most widely used visual features in retrieval.

Color histogram is the most commonly used color descriptor in content based retrieval research. Color histogram can be found in [5-8]. Michael and Dana proposed histogram intersection and the similarity measure for the color histogram in [6]. In [5], the authors proposed the use of Gaussian Mixture Vector Quantization (GMVQ) as a quantization method for color histogram generation and its experiment results have shown that it has better retrieval performance for color images than the conventional color histogram methods. Color histogram is easy to implement, but it does not take into account spatial information because it does not consider spatial distribution of color. In [9], Yining et al. proposed a compact color descriptor which could be indexed in a 3D color space. The descriptor consists of the representations of colors and their percentages in the region. The author claimed that this descriptor gave a more efficient indexing because of its low dimension compared with the traditional color histograms. Besides color histograms, color sets [10] and color moments [11] have also been applied to represent and retrieve images. Smith and Chang proposed a color set in [10], which was defined as a selection of colors from a quantized color space. Color set allows fast indexing and search because of its low dimension. Stricker and Orengo in [11] used the first three geometric moments to represent colors. One drawback of the color moment descriptor is that the average of all colors can be quite different from any of the original colors. Hence, like the traditional color histograms, given a color moment descriptor, it is impossible to recover the actual colors in the image.

3.2 Texture

Texture is defined as a local arrangement of image irradiances projected from a surface patch of perceptually homogeneous irradiances [12]. It is often used in medical images and satellite images. Tuceryan and Jain [13] identified five major categories of features for texture that

are statistical features, geometrical features, structural features, model-based features and features from signal processing. Chen and Li in [14] proposed texture-spectrum, a statistical way to describe texture features. The basic idea of using a texture spectrum is that a texture image can be represented as a set of essential small units. The statistics of all texture units over the entire image reveal a global texture feature. Wavelet transform as a kind of signal processing tool has been studied by many researchers to represent texture [15-18]. Chang and Kuo proposed a tree-structured wavelet transform to improve classification accuracy. In [15], Smith and Shih-Fu used the statistics extracted from wavelet subbands as a texture representation. In [16], texture features were modeled according to the marginal distribution of wavelet coefficients using generalized Gaussian distributions. Thyagarajan, Nguyen, and Persons combined the wavelet transform with a co-occurrence matrix to take advantage of both statistic based and transform based texture analysis [17].

3.3 Shape

Shape is a key attribute of image. Shape can be used to describe object's position, orientation and size. Therefore, it is required that shape representation should be invariant to translation, rotation and scale. In general, shape representation can be divided into two categories, boundary based and region based. The former uses only the outer boundary of the shape while the later uses the entire shape region. Fourier descriptor and moments are used for both categories. The most successful representation for shape is Fourier Descriptor (FD). FD is simple to derive. Coarse shape features or global shape features are captured from the lower order coefficients of FD and the finer shape features are captured from the higher order coefficients. Furthermore, FD is robust to noise because noise often coincides with very high frequency domains which are truncated out in FD. The FD's characters make it become a popular shape descriptor. However, FD cannot capture interior shape content which sometimes is important for shape description. For example, if two images have the same contour, but different interior, FD cannot discriminate their difference. 2-D Fourier transform in polar coordinates was employed for shape description in [19], which performed better than 1-D Fourier transform. Chuang and Kuo in [20] used a wavelet transform to describe the shape information, from which a wavelet descriptor is proposed. One advantage of wavelet descriptor is that it can provide global features at the coarser resolution levels and more detailed local features at the finer resolution levels. However, wavelet descriptor is not invariant because its transform coefficients are different for different starting points.

Moments are also popular to describe shape. Various types of moments have been used for moment based shape classification [21-23]. Hu introduced seven moments in

[21], which are invariant to translation, rotation and scaling. The advantages of moment descriptors are that it is easy to implement and different level moments are independent. However, it is difficult to associate higher order moments with physical interpretation, which makes the moment descriptors hard to understand and sensitive to even small geometric and photometric distortions [24].

Belongie, Malik and Puzicha introduced a new shape descriptor called shape context in [25]. 'Shape context' can be considered as a 3-D histogram of edge point location and distribution. 'Shape context' is robust to a number of transformations including translation, scale and rotation. The drawback of 'shape context' is that it relies much on its sample points. If sample points have a small error, the whole context will be totally incorrect. The challenge with shape based CBIR system is that the shape features need very accurate segmentation of image to detect the object or region boundary. Reliable segmentation is critical, without which shape representation is meaningless.

Color, texture, shape are all low level features of images. Now, more and more researches focus on the high level features of images and try to obtain the semantic features of images from low level features. A semantics-sensitive approach to content-based image retrieval has been proposed in [26, 27]. A semantic categorization (e.g. graph, photograph, indoor, outdoor, etc) for appropriate feature extraction has been used in [26]. The advantage of semantic categorization is that it improves retrieval accuracy and reduces image retrieval time. But the statement in [26] that similar semantics share similar visual features is not always true. Semantic categorization can also be used for human posture retrieval because human posture is a kind of high level features.

Human behaviour will use some technology in content based retrieval, but few work has been done on this specific work.

3.4 Similarity Measurement

Although measurement of feature similarity is application oriented, similarity measurement based on a statistical analysis has been dominant in content based retrieval.

Measuring the distance between histograms has been an active research stream for content-based retrieval when histogram is used for image representation. For content-based retrieval, histograms have mostly been used in conjunction with color features. But there is nothing against a conjunction with texture or shape properties. Michael and Dana [6] used the intersection distance

$$D(H(I), H(Q)) = \sum_{j=1}^n \min(H_j(I), H_j(Q)) \quad (1)$$

where $H(I)$ and $H(Q)$ are two histograms containing n bins each. In [28], a different approach was proposed. The distance between two histograms was defined in vector form as:

$$D_{hist} = (H(I) - H(Q))^T A (H(I) - H(Q)), \quad (2)$$

where D_{hist} is the distance of the histograms of two images, $H(I)$ and $H(Q)$ are histograms of the color vectors with K components and A is a $K \times K$ similarity matrix. This measurement considers the similarity between values in the feature space. Other commonly used distance functions for color histograms include Minkowski distance:

$$D(H(I), H(Q)) = \left[\sum_{j=1}^n |H_j(I) - H_j(Q)|^r \right]^{1/r} \quad (3)$$

The above measures do not take into account the similarity between different, but related bins of a histogram. In [29], the histogram was applied to color images by representing colors in the HSV color space and computing the channel moments separately, resulting in nine parameters, three moments for each of three color channels. The distance of color layout was also used for color similarity measure. As for color layout, a predefined grid color layout was used as a sample. For shape comparison, the match is based on transforms, moments, deformation, scale and etc. Because image can be represented as a wavelet or Fourier function, we can process the image through its coefficients. By truncating the coefficients below a threshold, images can be sparsely represented at the cost of loss of some detail. The set of remaining coefficients can be used as a feature vector for matching. One drawback of this approach is that it depends much on clear image segmentation. Moments, especially the low order moments, are often used for image retrieval. Scale space matching is based on progressively simplifying the contour through smoothing [30]. By comparing the signature of annihilated curvature zero crossings, two scale and rotation invariant shapes are matched. Salient features are used to capture the information in image in a limited number of salient points. Similarity between images can then be checked in several different ways. One method is to store all salient points from one image in a histogram on the basis of a few characteristics, and then the similarity is based on the presence of enough group-wise similar points [31]. The second method for similarity measurement of salient points is to concentrate only on the spatial relationships among the salient points sets. In point-by-point based methods as shown in [32] for shape comparison, shape similarity was studied, where maximum curvature points on the contour and the length between them were used to characterize the

object. Similarity at semantic level can be found in [33]. Knowledge based type abstraction hierarchy was used to access image data based on context and a user profile generated automatically from cluster analysis of the databases. In some cases, weights were used when features were fused together to calculate the similarity. The selection of weights corresponding to the features used affects the whole retrieval results.

3.5 Human Behavior Analysis

More and more efforts have been put on the research of human behavior analysis due to its promising applications in many areas such as visual surveillance, content-based image storage and retrieval, video conferencing, athletic performance analysis, virtual reality etc. A general frame for human behavior analysis is shown in Fig 2 [34].

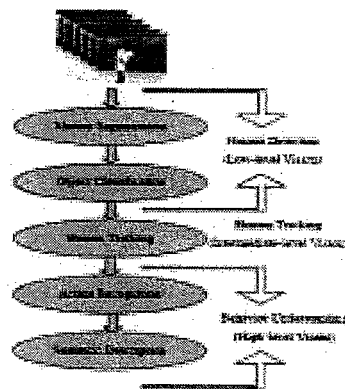


Fig 2 A general framework for human behavior analysis

Almost all methods for vision-based human behavior analysis starts with human detection. Human detection aims at segmenting regions corresponding to people from the rest of an image. It is a significant issue in a human behavior analysis system since the subsequent processes such as tracking and behavior understanding are greatly dependent on it. This process usually involves motion segmentation and object classification. After human detection, tracking will be applied. Tracking can be considered to establish coherent relations of image features between frames with respect to position, velocity, shape, texture, color etc. After successfully tracking moving humans from one frame to another in an image sequence, it is the turn to understand the behavior. Behavior understanding is to analyze and recognize human motion patterns, and to produce high-level description of actions and interactions. How to know some actions sequences can be one behavior is a problem needed to be solved for human behavior understanding. General approaches for behavior understanding are template matching [35] and

state-space [36]. For examples, HMM is an approach using for human behavior understanding, but it is not suitable for behavior retrieval because it requires beforehand models and models may be sensitive to noise.

Using natural languages to describe human behavior has received considerable attention. Its purpose is to reasonably choose a group of words or short expressions to represent the behaviors. Kojima et al. [37] proposed a method to generate natural language description of human behaviors appearing in real video sequence.

Human behavior retrieval has been researched in some specific areas, especially in the sport areas [38, 39]. The current human behavior retrieval techniques use other object's information to help determine human behavior. For example, the tennis ball's position, the line and net of a tennis court and the human body, are all needed to determine a player's action. The advantage of this method is that it can recognize human behavior accurately. But it requires more computation time and plentiful specific knowledge in the area.

4 Future Work on Human Behaviour Retrieval

Although human behaviour retrieval to some extent is similar to human behaviour analysis, the technology used in human behaviour analysis cannot be moved to human behaviour retrieval directly because the performance evaluation between these two areas is different. The metric used for human behaviour analysis is recognition rate. The higher recognition rate is the better performance will have. But for human behaviour retrieval, we should include not only the recognition rate as a metric, but also the recall rate and retrieval speed. It must find tradeoffs among these three metrics. Efficiently locating the video shots containing interested human behaviour based on given query data is the major aim of human behaviour retrieval.

The state-space methods shown in [36] can be used for human behaviour retrieval if we can find some solution to resist noise and reduce its computation time. Each state can overlap with others to some degree may help to solve this problem. But to what degree the states can overlap with each other should be carefully examined. The benefit using state-space method is that it can be combined with human behaviour analysis easily.

Since human behavior is composed of consecutive postures, it is possible to identify behavior by the consecutive postures. There are some benefits of using postures for behavior retrieval. The first is that it is not restricted in specific area, and it can be used in general areas. For example, it can be used to find some suspicious human behavior in an airport lobby. The second benefit is that it can find the targets very quickly. Using posture to retrieval human behavior, there is no guarantee that all of retrieved video are target videos. However, sacrificing

accuracy, as a return we can save large amount of time in processing. The third benefit is that postures have semantic meanings which are more understandable than the low level features.

5 Conclusions

This paper has given a brief review on the current techniques for human behavior analysis, video annotation, feature extraction and similarity measurement. All of these techniques are related with human behavior retrieval. This paper has also addressed possible future works on human behavior retrieval.

6 References

- [1] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: real-time surveillance of people and their activities," *Pattern Analysis and Machine Intelligence*, IEEE Transactions on, vol. 22, pp. 809-830, 2000.
- [2] G. Zhu, C. Xu, Q. Huang, and W. Gao, "Action Recognition in Broadcast Tennis Video," in *Pattern Recognition*, 2006. ICPR 2006. 18th International Conference on, 2006, pp. 251-254.
- [3] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos, "The Bayesian image retrieval system, PicHunter: theory, implementation, and psychophysical experiments," *Image Processing*, IEEE Transactions on, vol. 9, pp. 20-37, 2000.
- [4] A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases," *International Journal of Computer Vision*, vol. V18, pp. 233-254, 1996.
- [5] S. Jeong, C. S. Won, and R. M. Gray, "Image retrieval using color histograms generated by Gauss mixture vector quantization," *Computer Vision and Image Understanding*, vol. 94, pp. 44-66, 2004.
- [6] J. S. Michael and H. B. Dana, "Color indexing," *International Journal of Computer Vision*, vol. V7, pp. 11-32, 1991.
- [7] M. Ortega, Y. Rui, K. Chakrabarti, K. Poraew, S. Mehrotra, and T. S. Huang, "Supporting ranked Boolean similarity queries in MARS," *Knowledge and Data Engineering*, IEEE Transactions on, vol. 10, pp. 905-925, 1998.
- [8] W. Jia, H. Zhang, X. He, and Q. Wu, "A Comparison on Histogram Based Image Matching Methods," 2006, pp. 97-97.

- [9] D. Yining, B. S. Manjunath, C. Kenney, M. S. Moore, and H. Shin, "An efficient color representation for image retrieval," *Image Processing, IEEE Transactions on*, vol. 10, pp. 140-147, 2001.
- [10] J. R. Smith and S. F. Chang, "Single color extraction and image query," 1995, pp. 528-531 vol.3.
- [11] M. Stricker and M. Orengo, "Similarity of Color Images," in *SPIE Storage and Retrieval for Image and Video* vol. 2420, 1995, pp. 381-392.
- [12] A. C. Bovik, M. Clark, and W. S. Geisler, "Multichannel texture analysis using localized spatial filters," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 12, pp. 55-73, 1990.
- [13] M. Tuceryan and A. K. Jain, "Texture Analysis," in *The Handbook of Pattern Recognition and Computer Vision* (2nd Edition), by C. H. Chen, L. F. Pau, P. S. P. Wang (eds.), 2, Ed.: World Scientific Publishing Co., 1998, pp. 207-248.
- [14] H. Dong-chen and W. Li, "Texture Unit, Texture Spectrum, And Texture Analysis," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 28, pp. 509-512, 1990.
- [15] J. R. Smith and C. Shih-Fu, "Automated binary texture feature sets for image retrieval," 1996, pp. 2239-2242 vol. 4.
- [16] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance," *Image Processing, IEEE Transactions on*, vol. 11, pp. 146-158, 2002.
- [17] K. S. Thyagarajan, T. Nguyen, and C. E. Persons, "A maximum likelihood approach to texture classification using wavelet transform," 1994, pp. 640-644 vol.2.
- [18] T. Chang and C. C. J. Kuo, "Texture analysis and classification with tree-structured wavelet transform," *Image Processing, IEEE Transactions on*, vol. 2, pp. 429-441, 1993.
- [19] Z. Dengsheng and L. Guojun, "Generic Fourier descriptor for shape-based image retrieval," 2002, pp. 425-428 vol.1.
- [20] G. C. H. Chuang and C. C. J. Kuo, "Wavelet descriptor of planar curves: theory and applications," *Image Processing, IEEE Transactions on*, vol. 5, pp. 56-70, 1996.
- [21] H. Ming-Kuei, "Visual pattern recognition by moment invariants," *Information Theory, IEEE Transactions on*, vol. 8, pp. 179-187, 1962.
- [22] H.-K. Kim and J.-D. Kim, "Region-based shape descriptor invariant to rotation, scale and translation," *Signal Processing: Image Communication*, vol. 16, pp. 87-93, 2000.
- [23] A. Khotanzad and Y. H. Hong, "Invariant image recognition by Zernike moments," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 12, pp. 489-497, 1990.
- [24] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, pp. 1615-1630, 2005.
- [25] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, pp. 509-522, 2002.
- [26] J. Z. Wang, L. Jia, and G. Wiederhold, "SIMPLiCity: semantics-sensitive integrated matching for picture libraries," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, pp. 947-963, 2001.
- [27] J. Feng, L. Mingjing, Z. Hong-Jiang, and Z. Bo, "An efficient and effective region-based image retrieval framework," *Image Processing, IEEE Transactions on*, vol. 13, pp. 699-709, 2004.
- [28] J. Hafner, H. S. Sawhney, W. Equitz, M. Flickner, and W. Niblack, "Efficient color histogram indexing for quadratic form distance functions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 17, pp. 729-736, 1995.
- [29] M. Stricker and M. Orengo, "Similarity of Color Images," *Storage and Retrieval for Image and Video* vol. 2420, pp. 381-392, 1995.
- [30] F. Mokhtarian, "Silhouette-based isolated object recognition through curvature scale space," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 17, pp. 539-544, 1995.
- [31] T. Gevers and A. W. M. Smeulders, "PicToSeek: combining color and shape invariant features for image retrieval," *Image Processing, IEEE Transactions on*, vol. 9, pp. 102-119, 2000.
- [32] J. Linhui and L. Kitchen, "Object-based image similarity computation using inductive learning of contour-segment relations," *Image Processing, IEEE Transactions on*, vol. 9, pp. 80-87, 2000.
- [33] H. Chih-Cheng, W. W. Chu, and R. K. Taira, "A knowledge-based approach for retrieving images by

content," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 8, pp. 522-532, 1996.

[34] L. Wang, W. Hu, and T. Tan, "Recent Developments in Human Motion Analysis," *Pattern Recognition*, vol. 36, pp. 585-601, 2003.

[35] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, pp. 257-267, 2001.

[36] A. Galata, N. Johnson, and D. Hogg, "Learning Variable Length Markov Models of Behaviour," *Computer Vision and Image Understanding*, vol. 81, pp. 398-413, 2001.

[37] A. Kojima, M. Izumi, T. Tamura, and K. Fukunaga, "Generating natural language description of human behavior from video images," 2000, pp. 728-731 vol.4.

[38] H. Miyamori and S. I. Iisaku, "Video annotation for content-based retrieval using human behavior analysis and domain knowledge," 2000, pp. 320-325.

[39] G. Sudhir, J. C. M. Lee, and A. K. Jain, "Automatic classification of tennis video for high-level content-based retrieval," 1998, pp. 81-90.