ACCV2002: The 5th Asian Conference on Computer Vision, 23--25 January 2002, Melbourne, Australia

A Novel Approach to Video Indexing and Retrieval Using Motion Activity

Lujun Yuan[1] Wen Gao[1, 2] Wei Zeng [2]
[1](*Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080)*
[2](*Department of Computer Science and Engineering, Harbin Institute of Technology,150001)*
ljyuan@ict.ac.cn wgao@ict.ac.cn wzeng@ict.ac.cn

Abstract

Motion activity description plays an important role on video retrieval, which quantitatively provides the intensive degree of scene motion from the point of human perception. An approach based on local activity histogram (LAH) is proposed to describe motion activity. Firstly, the motion vector field is abstracted from video. Then, the motion vector field is transformed to the intensity tag image (ITI) by quantization operation. Based on the idea of spatial template, a small window scans the ITI and counts the number of different template with varied intensity tags. As a result, a local activity histogram is computed to represent the distribution of intensity tags. By combining intensity of motion vector field and LAH, a descriptor for motion activity is obtained to represent video motion activity. Furthermore, our method supports region-based motion activity description. Experiment result shows that our method has a better retrieval precision and flexibility on motion-based retrieval.

1. Introduction

Content-based multimedia retrieval (CBMR) is an active research area in recent years. CBMR wants to provide efficient methods to describe and classify the multimedia documents and make the multimedia contents be accessed more easily. By the efforts of researchers, many multimedia retrieval systems have been developed now. IBM has developed the famous QBIC image/video retrieval system [1] and integrated the techniques to their product DB2 database system. MIT Media Lab has developed the PHOTOBOOK image retrieval system by using interactive learning concept [2]. In UIUC (Univ. Of Illinois at Urbana Campus, USA), Prof. Thomas Huang's research group has developed MASRS image/video retrieval system based on relevance feedback technique [3]. Prof. Shi-fu Chang etc. have developed the object oriented video retrieval system Visual SEEK [4] in the Columbia University. At the same time, MPEG started up to draft out MPEG-7standard in 1996. MPEG aims at providing a set of standard for description audio-visual contents in multimedia environments [5].

To characterize the contents of image/video and allow efficient indexing and retrieval, researchers have suggested many low-level visual features such as color, texture, shape, motion etc [6,7]. Above all, the motion feature is the main different characteristic between image and video, which provides the temporal information of video. Many techniques have been explored by researchers to extract the motion features:

1) Analyze the camera motion and index video with camera motion characteristics.
2) Compute motion model parameters and index video with these model parameters.
3) Segment the object from the video and index video with the object motion trajectory.

The aforementioned methods are based on the point of computer vision. However, the motion activity feature has been proposed to annotate video segments to capture the intuitive notion of 'intensity of action' or 'pace of action' in a video segment [9,10,11]. For example, 'goal scoring in a soccer match', ' a high speed car chase' are high activity video scenes, on the contrary, 'news reader shot',' an interview scene' are low activity scenes. Based on the concept of motion activity, a similarity of video can be defined by comparing video's motion activity. Some related work is as follows:

Ajay Divakaran etc. [9] proposed a region-based method for description spatial distribution of motion activity in compressed video sequences. They use a histogram of different areas regions with different motion activity to describe spatial distribution of motion activity. The main problem of their approach is the number of bins in the histogram must be determined manually. If the number of video shots is large, it is very difficult to determine the number of bins manually.

Kadir A. Peker etc. [10] use the average block-based motion vector magnitude and the average temporal derivation of motion vectors to describe the motion activity. They aim to establish certain connections between low-level feature and high-level semantic characterization of the video segments. The drawback of their method is that they do not utilize the spatial distribution information of motion vectors.

MPEG-7 eXperimental Model (XM) [11] uses the intensity of motion vector and the run length information to describe the motion activity. In this method, the intensity and the spatial distribution characteristics of motion vectors provided the good description for the motion activity. However, run length characteristic is limited in describing the spatial distribution of motion vectors while to compute motion activity in different size region. Therefore, it is not flexible in the region-based retrieval.

In this paper, we propose an approach that uses the intensity of motion and the local activity histogram to describe motion activity. Firstly, we compute the motion vectors between two adjacent video frames; secondly, we use the average and the variance of the motion vector magnitude to represent the intensity of motion; at last, we quantize motion vectors to form a intensity tag image (ITI) and scan ITI using a specific spatial template to obtain the local activity histogram, which can well describe the spatial distribution of motion activity. Experiment results show that our method enables accurate and flexible video retrieval.

Rest of the paper is organized as follows. In section 2, the video indexing and retrieval framework by motion activity is proposed firstly, and then we describe how to compute an activity feature vector to indexing video shots, at last we propose a new shot-based retrieval scheme. Experimental results are given in section 3. Finally, section 4 presents the conclusions and future work.

2. Indexing and retrieval using motion activity

The motion activity characteristic captures the intuitive notion of ‘intensity of action’ or ‘pace of action’ in a video segment. In order to describe the video activity quantitatively, firstly we compute the motion vector field between two adjacent video frames, and then compute a feature vector to represent motion activity. At last, we define the activity similarity measure to index and retrieve the video segments. Figure 1 shows the video indexing and retrieval framework using motion activity.

2.1. Computing motion vectors

In our work, motion vectors are computed using an exhaustive-search block-matching algorithm. Although some motion vectors do not represent the “real motion”, the spurious motion vectors cannot spoil the result in the practical application. The base idea of block matching is as follows:

1) Each frame is divided into $N_1 \times N_2$ blocks.

2) Searching the reference frame for the location of the best-matching block in terms of the certain matching criterion. The search is usually limited to a $(N_1+M_1)\times(N_2+M_2)$ region called the search window for computational reasons. The most commonly used matching criterion is MAD (mean absolute distortion), which is given by:

$$MAD=\frac{1}{N_1N_2}\sum_{i=0}^{N_1}\sum_{j=0}^{N_2}\left|S_1(x+i,y+j)-S_2(x+i,y+j)\right|$$

Where S_1 denotes the block in the current frame, and S_2 denotes the candidate block in the reference frame.

If the video segments are MPEG video streams, we can extract the motion vectors from the video streams directly without the time-consuming block matching procedure.

2.2. Frame-based motion activity

The 2-D motion vector field consists of all motion vectors between two adjacent video frames. Generally, each motion vector has the intensity characteristic and the directional characteristic. Because the activity feature represents the “intensity of action”, we estimate motion activity only using the intensity of the motion vector. At the same time, we use the local histogram to describe the distribution of motion activity. Hence the activity feature can be defined as follows:

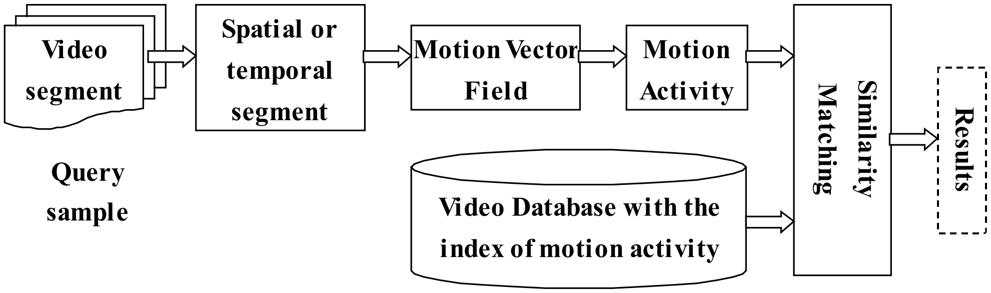


Figure 1. The retrieval framework based motion activity

$$F = (\bar{f}, f_{\text{var}}, h_1, h_2, h_3, \cdots)$$

Where $\bar{f}$ denotes the average intensity, f_{var} denotes the variance of the intensity, and $h_1, h_2, h_3, \cdots$ are the bins of the local activity histogram.

2.2.1. Intensity characteristics. Consider *m* × *n* motion vectors in the 2-D motion vector field, $\vec{v}_{i,j}$ represents the motion vector at the spatial position (*i, j*). The intensity characteristics of motion activity $\bar{f}$ and f_{var} are given by:

$$\bar{f} = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} \left| \vec{v}_{i,j} \right|$$

$$f_{\text{var}} = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\left| \vec{v}_{i,j} \right| - \bar{f} \right)^2$$

2.2.2. Local activity histogram. The local activity histogram (LAH) is used to describe the spatial distribution of motion activity. We can construct LAH of the 2-D motion vector field using the following method:

1) Find the intensity tag image (ITI) by uniformly quantizing the motion vector magnitude. Each quantization level is represented with an intensity tag $R(K)$.

2) Scan ITI using a specific spatial template structure (e.g. 4-neighborhood template structure), and count the relative frequency of occurrence of various templates, then LAH is defined as follows:

$$H = (h_1, h_2, \cdots, h_N)$$

$$h_i = \frac{num\ (\Omega_i)}{\sum_{j=1}^{N} num\ (\Omega_J)}$$

Where Ω_i is the *i*th local activity template in the template space $\Re$, *num*(.) is the count function. The local activity template Ω_i is given by:

$$\Omega_i = \{\{R(1), R(2) \ldots R(L)\}, \\ \{n^{\Re}_{R(1)}, n^{\Re}_{R(2)}, \cdots, n^{\Re}_{R(L)}\}\}$$

Where $R(i)$ denotes the strength tag, $n^{\Re}_{R(i)}$ denotes the number of the intensity tag $R(i)$ in this template. If the spatial template structure is *n-neighborhood* and the number of intensity tags is *r*, the number of bins in the LAH is:

$$N = C_{n+r-1}^{r-1}$$

Here we provide an example to illustrate. Consider a 4-neighborhood spatial template structure and a 2-level quantization operation, the intensity tags are defined as follows:

$$\begin{cases} if \quad f(i,j) < \bar{f} & R(0) = 0 \\ else & R(1) = 1 \end{cases}$$

$$Where: f(i,j) = \left| \vec{v}_{i,j} \right|$$

Therefore, there are 5 local activity templates, which shown in Figure 2.

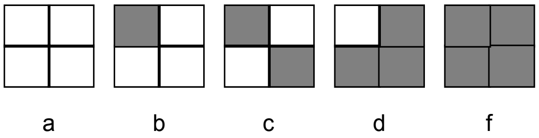


Figure 2. Local activity templates

In Figure 2, the deep color block represents the low activity tag $R(0)$. Template (a) and (e) are very high activity region and very low activity region respectively. Template (b), (c), (d) represent the motion boundary region. As a result, a feature vector is defined to describe motion activity as follows:

$$F = (\bar{f}, f_{\text{var}}, h_1, h_2, h_3, h_4, h_5)$$

2.3. Shot-based motion activity

Obtaining all frame-based motion activity feature vector of one video shot, we can extract the activity feature for the whole video shot. The activity feature of the shot can be computed using one of the following methods:

1) Choose one from all 2-D motion vector fields of the shot and compute its activity feature vector to represent the activity feature of the shot.

2) Compute the average of all activity feature vectors to represent the activity feature of the shot.

3) Compute the average of all activity feature vectors, and then choose the activity feature vector that is the most similar to the average vector to represent the activity feature of the shot.

4) Compute the average of all motion vector fields in the shot, and then extract the activity feature vector of the average motion vector field to represent the activity feature of the shot.

After obtaining the motion activity feature vector of the shot, we can define the distance of two video shots as follows:

$$\text{Distance}(q, p) = v_1 \times \exp(v_2)$$

Where:

$$v_2 = \left|\bar{f}_q - \bar{f}_p\right|$$

$$v_1 = \frac{w}{f_{\text{var}}^q}\left|f_{\text{var}}^q - f_{\text{var}}^p\right| + \sum_{i=1}^{N} \frac{w}{m_i^q}\left|m_i^q - m_u^p\right|$$

$$w = f_{\text{var}}^q + \sum_{i=1}^{N} m_i^q$$

In this equation, the exponential term represents the motion activity difference of video shots, for the human visual system is sensitive to the motion intensity change, and the similarity of video shots is the exponential relationship with the motion intensity difference, and the term v_1 is the weight sum of the other feature components.

2.4. Region-based motion activity

In some applications, users may only concern the activity in a specific region in a video frame. They want to find shots with the similar activity in the certain region. Our method can easily extend to support this region-base retrieval.

Let *R* represents the region of interest, the template structure is 4-neighborhood and the number of intensity tags is 2, the activity feature vector for the region-base retrieval is given by:

$$F^R = (\bar{f}^R, f_{\text{var}}{}^R, h_1{}^R, h_2{}^R, h_3{}^R, h_4{}^R, h_5{}^R)$$

Where:

$\bar{f}^R = \frac{1}{c}\sum\left|v_{(i,j)}\right|$ is the average intensity in the region *R* (*c* is the number of motion vectors in the region *R).*

$f_{\text{var}}^R = \frac{1}{c}\sum(\left|v_{i,j}\right| - \bar{f}^R)^2$ is the variance of motion vector magnitudes in the region *R*.

h_i^R is the *i*th bin of the local activity histogram of the region R.

In this case, the distance of two video shots is defined as follows:

$$\text{Dist}(q, p, R) = v_1^R \times \exp(v_2^R)$$

Where v_1^R and v_2^R have the same meaning as the v_1 and v_2 in the equation of section 2.3.

2.5. Shot-based retrieval

Because human being pays different attention to different region when he/she watching a video sequence, we divide the full frame to several regions and extract the activity feature vectors for each region, and then compute the region-based distance. Finally we combine these distances to compute the shot distance. The new shot distance is given as follows:

$$\text{Dist}(q, p) = \sum_i w_i Dist(q, p, R_i)$$

Where w_i is the visual weight in the region R_i.

3. Experimental results

Our video database consists of 27-minute NBA basketball match, 17-minute tennis match, 13-minute soccer match, 17-minute sports news and 37-minute CCTV news, which have total 1064 video shots. In order to test the performance of our proposed algorithm, we annotate 37 news-anchorperson shots, 145 tracking shots, 51 basketball shots, 45 soccer shots and 53 tennis shots.

3.1. Experiment 1

In the region-based retrieval experiments, our test video sequences consist of 14-minute CCTV news with 81 video shots, 27-minute basketball match with 83 shots and 13-minute soccer match with 121 shots. The results are shown in Figure 3 (the regions of interest are marked with white rectangles).

3.2. Experiment 2

We use all the 1064 video shots in the shot-based retrieval experiment. Two metrics, precision and recall, are used to evaluate our algorithm. These two metrics are defined as follows:

$$\text{Re}\,call = \frac{Detects}{Detects + MD.'s}$$

$$\text{Pr}\,ecision = \frac{Detects}{Detects + FA.'s}$$

Where *Detects* is the correct shots in the best match shots, *FA's* is the false shots in the best match shots; *MD's* is the number of missed detections.

In our experiments, we compute the average precision and recall for each annotated shot class. In order to give a comparison, we also do the same experiments using the algorithm provided by MPEG-7. The results are shown in Table 1 and Figure 4.

From experimental results, we find our algorithm has perfect performance in the video retrieval where the motion regions are uniformly distributed, such as news anchorperson shots and shooting shots in the soccer match. The anchorperson shots are typical low activity video shots with small-range motions of the mouth and head. Shooting shots are typical high activity video shots with the high-speed camera motion. In the tracking shots with large-area moving foreground objects, our result is similar to MPEG-7's.

4. Conclusions

In this paper, we propose a new approach for video retrieval based on motion activity. Our method can support both the shot-based retrieval and region-based retrieval in the same framework. We use the local activity histogram to describe the spatial distribution of motion vectors. By combining intensity characteristics of motion vectors and LAH, a descriptor for motion activity is obtained to represent the motion activity feature. Experimental results show that our method has a better retrieval precision and flexibility in the video retrieval.

In the future work, we will combine motion activity and human's perceptive model to form more appropriate video retrieval for users.

5. References

[1] M.Flickner et al., "Query by Image and Video Content: The QBIC System", Computer, vol. 28, no.9, 1995, pp. 23-32

[2] Thomas P. Minka and Rosalind W. Picard, "Interactive learning using a 'society of models' ", In Proceeding of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'96), 1996, pp. 447-452

[3] Yong Rui, Thomas S. Huang, Michael Ortega, and Shard Mehrotra, "Relevance Feedback: A Powerful Tool in Interactive Content-Based Image Retrieval", IEEE Tran on Circuits and Systems for Video Technology, Vol. 8, No.5, 1998, pp. 644-655

[4] S.F Chang, W. Chen, H.J. Meng, H. Sundaram, and D. Zhong, "A Fully Automatic Content-based Video Search Engine Supporting Multi-Object Spatio-temporal Queries", IEEE Tran on Circuits and Systems for Video Technology, Vol. 8, No.5, Sept.1998, pp. 602-615

[5] MPEG-7 Overview (version 2.0), ISO/IEC JTC1/SC29/WG11, document no. N3349, Noordwijkerhout (The Netherlands), March 2000

[6] Alberto Del Bimbo, "Visual Information Retrieval", Morgan Kaufmann Publishers, USA, 1999

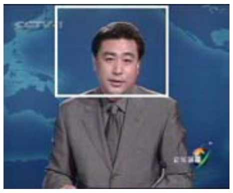
[7] R. Brunelli, O. Mich and C. M. Modena, “A Survey on Video Indexing”, Journal of Visual Communication and Image Representation, No.2, 1999

[8] Hsu P.R., Harashima H. “Detecting Scene Changes and Activities in Video Databases”， In Proceeding of IEEE International Conference on Acoustic, Speech, and Signal Processing, Adelaide, 1994

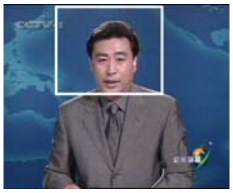
[9] Ajay Divakaran, Kadir Peker, Huifang Sun, "A Region Based Descriptor for Spatial Distribution of Motion Activity for Compressed Video", In Proceeding of IEEE International Conference of Image Processing (ICIP2000), 2000

[10] Kadir A. Peker, A. Aydin Alatan, Ali N. Akansu, "Low-level Motion Activity Features for Semantic Characterization of Video", In Proceeding of IEEE International Conference on Multimedia and Expo. (ICME2000), 2000

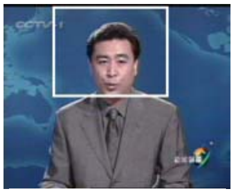
[11] MPEG-7 Visual Part of the eXperimentation Model Version 6.0, ISO/IEC JTC1/SC29/WG11, document no. N3398, Geneva, June 2000



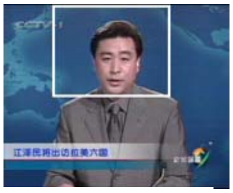
Query (id=12)



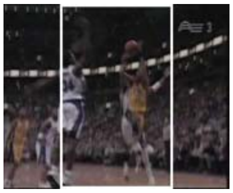
Match 1 (id=1)
Distance=6.423



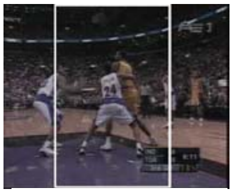
Match 2 (id=76)
Distance=9.294



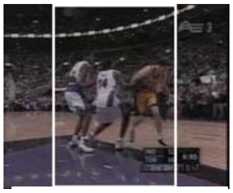
Match 3 (id=22)
Distance=9.598



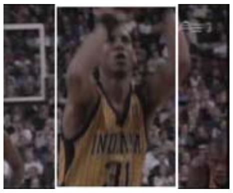
Query (id=65)



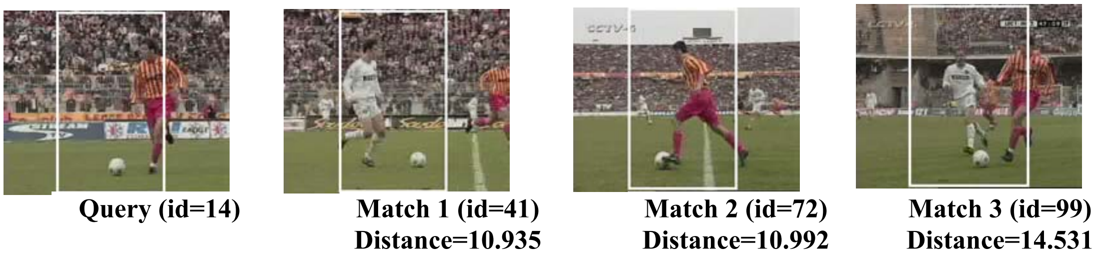
Match 1 (id=8)
Distance=12.216



Match 2 (id=47)
Distance=16.769



Match 3 (id=68)
Distance=17.541


Query (id=14)

Match 1 (id=41)
Distance=10.935

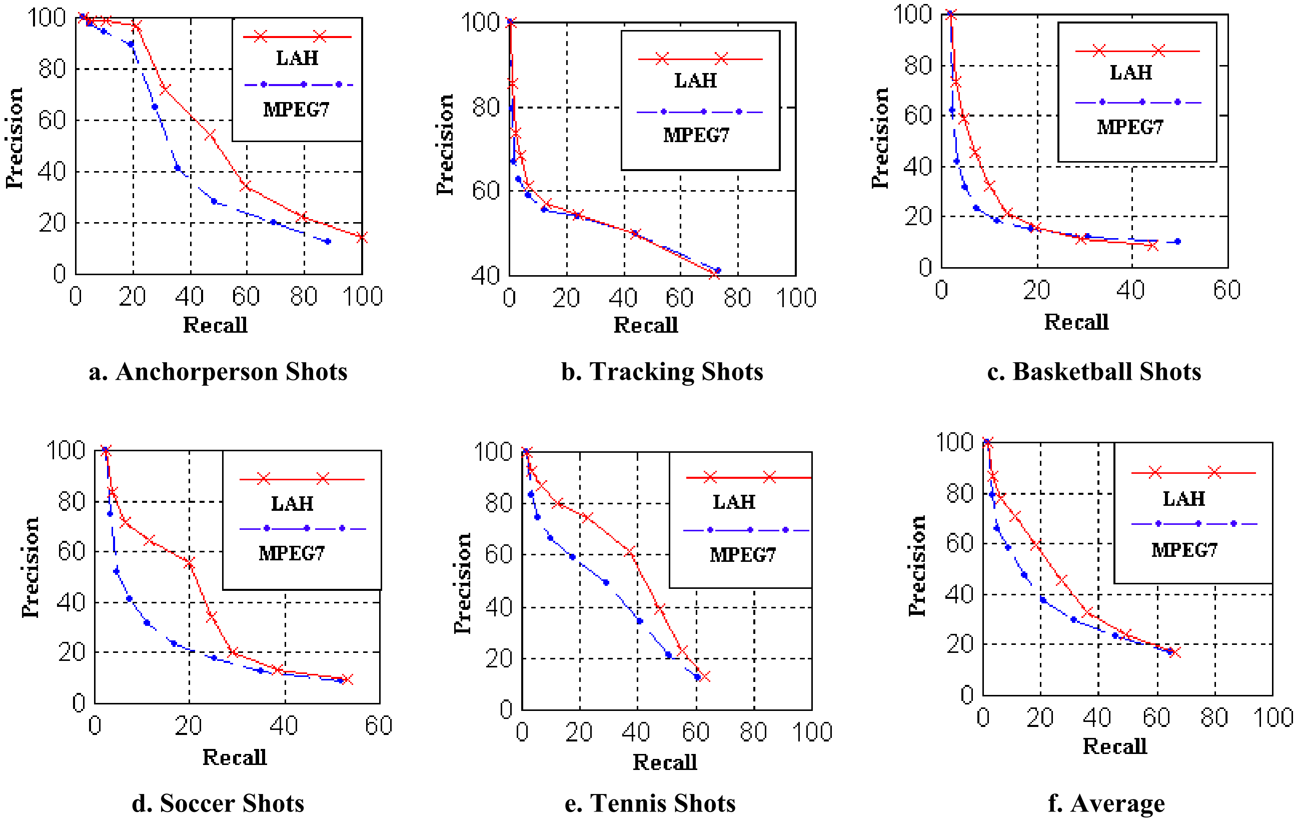
Match 2 (id=72)
Distance=10.992

Match 3 (id=99)
Distance=14.531

Figure 3. Results of the region-base retrieval

Table 1. The average precision and recall of the top 32 video shots

Samples	Detects / Precision		Ground-truth Video Shots	Recall	
	LAH	MPEG-7		LAH	MPEG-7
Anchorperson Shots	17.38 / 54.3%	13.30 /41.6 %	37	47.0 %	35.9%
Tracking Shots	18.28 / 57.1%	17.75 / 55.5%	145	12.6 %	12.2%
Basketball Shots	6.98 / 21.8%	5.90 /18.4%	51	13.7%	11.6%
Soccer Shots	11.00 / 34.3%	7.58 / 23.7%	45	24.4 %	16.8%
Tennis Shots	19.68 / 61.5%	15.60 / 48.8%	53	37.1%	29.4%
Average	14.66 / 45.8%	12.03 / 37.6%		27.0%	21.2%



a. Anchorperson Shots

b. Tracking Shots

c. Basketball Shots

d. Soccer Shots

e. Tennis Shots

f. Average

Figure 4. Recall-Precision graphs of the three shot classes