Video Concept Detection Using Support Vector Machines - TRECVID 2007 Evaluations

Zheng-Jun Zha[1], Yuan Liu[1], Tao Mei [2], Xian-Sheng Hua [2]

[1] University of Science and Technology of China
{zzjun; susie}@mail.ustc.edu.cn

[2] Microsoft Research Asia
{tmei; xshua}@microsoft.com

Abstract. This report describes video concept detection using Support Vector Machine (SVM) over TRECVID 2007 corpus. We perform the experiments on low-level features extraction, data preparation and classification procedure. Through analyzing the characteristics of the TRECVID 2007 data set, we mainly focus on data preparation for training concept detectors, as well as the preparation of auxiliary training data by using TRECVID 2005 data.

Keywords: concept detection, high-level feature extraction, support vector machines

1 Introduction

With rapid advances in storage devices, networks, and compression techniques, large-scale video data become available to more and more average users. To facilitate browsing and searching these data, it has been a common theme to develop automatic analysis techniques for deriving metadata from videos which describe the video content at both syntactic and semantic levels. With the help of these metadata, tools and systems for video summarization, retrieval, search, delivery, and manipulation can be effectively created. The main challenge is to understand video content by bridging the semantic gap between the video signals on the one hand and the visual content interpretation by humans on the other.

This semantic information is also called "high-level" features for describing, indexing and searching video content. Owing to the importance of video understanding, TREC Video Retrieval Evaluation (TRECVID) [1], organized by the National Institute of Standards and Technology, in its 7th year of evaluation on high-level feature task provides hundreds of hours of broadcast news videos and other raw videos. It focuses its efforts to promote progress in content-based retrieval from video via an open, metrics-based evaluation, based on the common video datasets and a standard set of queries.

In this report, we focus on understanding video content from the perspective of concept detection (also referred to as high-level feature extraction [1] or video

This work was performed when the first two authors were visiting Microsoft Research Asia as research interns.

annotation) over the benchmark data set provided by TRECVID 2007. In TRECVID 2007, we took part in two main tasks, i.e., high-level feature extraction and automatic search task. Both tasks are heavily relied on the performance of concept detection. In this paper, we mainly present the baseline method, which is based on a variety of low-level visual features and the employment of Support Vector Machine (SVM) as it proved to be the most effective machine learning techniques for concept detection.

Despite the simple features extraction and classification methods used for the concept detectors, there are some important issues deserving a certain attention which were not mentioned in most existing papers. Specifically, we will focus on data preparation for training concept detectors, as well as the preparation of auxiliary training data by using TRECVID 2005 data. After four years on broadcast news, TRECVID 2007 tests its tasks on new, related, but different video genre taken from a real archive. Concerned with the changes in the data characteristics and highly unbalanced distribution of samples, data partition and auxiliary data selection deserve a special attention as they will significantly affect the performance of concept detectors. We will first present nine types of low-level visual features, and then discuss the details of data preparation, classification procedure, computational complexity and performance evaluation.

2 Low-level Feature extraction

We extracted nine types of low-level global feature for each key-frame, which is provided by NIST [1]. In total, there are 1181 dimensional global features. Table 1 lists the detailed information about these features.

Table 1. Low-level global features used for concept detection.

Feature Name	Dimension	Description
AutoCorrelogram	144	Based on 36 bin color histogram and 4 different distance k, i.e., k = 1, 3, 5, 7.
ColorMoment3-by-3	81	Based on 3 by 3 grid division of images in Lab color space
ColorMoment5-by-5	225	Based on 5 by 5 grid division of images in Lab color space
ColorMoment7-by-7	441	Based on 7 by 7 grid division of images in Lab color space
Co-occurrence Texture	16	The same feature as in [6]
Edge Distribution Histogram	75	The same feature as in [6]
Face	7	Consisting of the number of faces, the ratio of face area, the position of the largest face
HSV Color Histogram	64	Global color represented as a 64-dimensional histogram in HSV color space [6]
Wavelet PWT&TWT Texture	128	The same feature as in [6]

3 High-level Feature Extraction

3.1 Data Preparation

3.1.1 Data Analysis

The TRECVID 2007 data set [1] is composed of 219 video clips, separated into two groups by NIST [2], i.e., the development set and test set. Specifically, the development set contains 110 video clips, consisting of 21,532 key-frames. These 36 concepts are manually annotated over these key-frames and the ground truth is provided by NIST. While the test set is composed of 109 video clips and is used for system testing. We extracted 30,661 key-frames from the test set.

Through analyzing the development set, we found that the numbers of positive and negative samples are highly unbalanced, especially for some concepts. Figure 1 illustrates the distribution of positive examples. The labels above the bins denote the number of positive samples of the corresponding concepts. The detailed sample distribution of TRECVID 2007 training data set can be found in Appendix A. For some concepts such as "Airplane," "Bus," and "Flag-US", the positive samples are extremely less than the negative ones, which will lead to the difficulties in model construction.

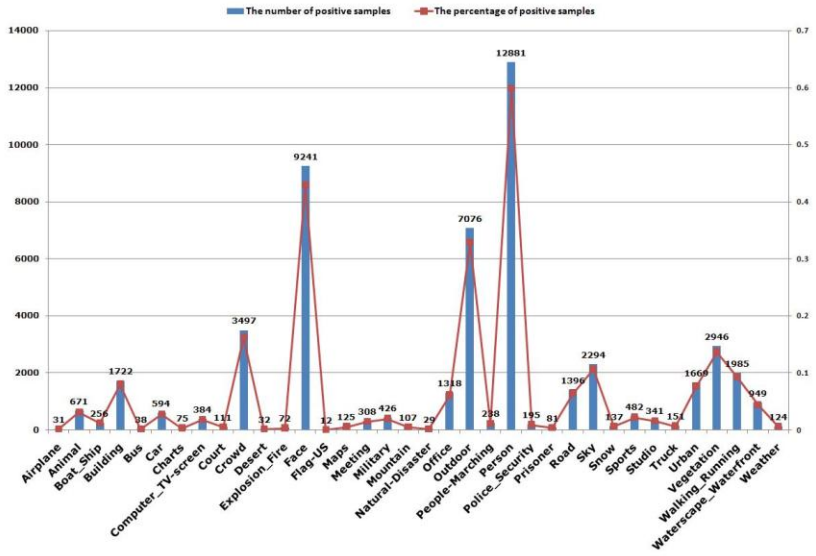

Figure 1: The distribution of positive/negative concepts in TRECVID 2007 data.

3.1.2 Data Partitioning

Since only the development set contains the annotations, we further partitioned it into four internal partitions for constructing and validating learning-based models. Table

2 gives the number of videos and key-frames in each subset. The "Train" subset is used to learn the concept detection models (i.e., detectors), while the "Validation" subset is used for model parameter selection. Moreover, we perform multimodal fusion over the "Fusion" subset and test the obtained concept detectors on the "Selection" subset. The number of the positive samples of each concept in each partition is listed in Table 3.

Table 2: Data partition of development set used to build concept detectors.

Dataset	Partition	# of Videos	# of Key-frames
TRECVID 2007 Development set	Train	77	14419
	Validation	11	2331
	Fusion	11	2298
	Selection	11	2484

Table 3: The number of positive samples in each partition.

Concept	Train	Validation	Fusion	Selection
Airplane	9	18	2	2
Animal	506	43	79	43
Boat_Ship	149	56	21	30
Building	992	225	244	261
Bus	29	3	3	3
Car	407	50	82	55
Charts	42	15	12	6
Computer_TV-screen	285	48	26	25
Court	0	0	0	111
Crowd	2330	420	451	296
Desert	4	22	3	3
Explosion_Fire	53	5	9	5
Face	5957	996	1023	1265
Flag-US	12	0	0	0
Maps	63	37	3	22
Meeting	188	38	39	43
Military	141	129	62	94
Mountain	47	23	16	21
Natural-Disaster	6	15	5	3
Office	766	103	127	322
Outdoor	4410	828	1034	804
People-Marching	129	36	18	55
Person	8483	1397	1327	1674
Police_Security	131	10	22	32
Prisoner	23	0	46	12
Road	915	120	219	142
Sky	1433	301	328	232
Snow	86	23	22	6
Sports	218	185	45	34
Studio	128	19	19	175
Truck	81	39	8	23
Urban	1044	227	245	153
Vegetation	1809	390	400	347

Walking_Running	1269	186	297	233
Waterscape_Waterfront	622	132	77	118
Weather	62	21	12	29

3.2 Support Vector Machine

We posed the multi-label video annotation task into binary classification problem. Support Vector Machine (SVM) [3] is adopted as the baseline due to its satisfactory effectiveness in concept detection. Figure 2 shows the pipeline for baseline detectors construction. As described in section 3.1, we extracted nine types of low-level visual features, among which eight features were utilized to build SVM classifiers. SVM classifiers are trained individually over each of the eight feature spaces. The SVMs are implemented using LIBSVM (Version 2.8) [3]. Then we perform modality fusion to obtain the detection result. The fusion methods include **Linear**, **Average**, and **Max**. We will give more details about the feature normalization, the auxiliary data selection, the model parameters selection, and the modality fusion in the next.

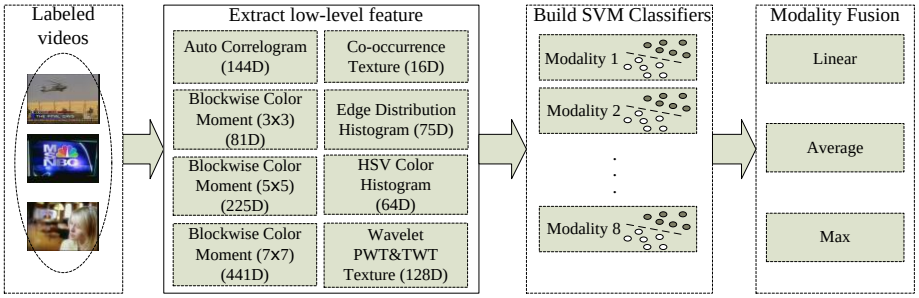

Figure 2: The Baseline detectors construction pipeline.

3.2.1 Feature Normalization

SVMs work well when features are roughly in the same range. Here, we normalize the features using statistical normalization [4] which shifts the feature distribution to zero mean (i.e., μ=0) and unified standard deviation (σ). For N feature vectors $\{x_1, x_2 \cdots, x_N\}$ in which x_i is an m-dimensional feature vector $[x_{i1}, x_{i2}, \cdots, x_{im}]^T$, we calculate the mean vector ($\boldsymbol{\mu}$) and the standard deviation vector ($\boldsymbol{\sigma}$) as follow.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i , \quad \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{1}$$

The features are then normalized through the following Equation,

$$x^* = \frac{(x - \mu)}{\sigma} \tag{2}$$

where $\boldsymbol{x}^*$ is the normalized feature. The division operation is applied to each component of the feature vector respectively.

3.2.2 Auxiliary Data Selection

As shown in Table 3, the positive samples of some concepts are highly sparse, such as "Airplane," "Desert," and "Natural-Disaster." This leads to the difficulties in building concept detectors. To address this problem, we manually selected some positive samples from TRECVID 2005 development set and added them into the TRECVID 2007 "Train" subset. Then, we constructed the baseline models on the expanded "Train" subset. They are named as "TRECV0705Model." Table 4 shows the concepts for which we select some positive samples from TRECVID 2005 development set. It also lists the number of the auxiliary positive samples. In addition,

- for "Flag-US," we also added some TRECVID 2005 positive samples into the "Validation" and "Fusion" subset since there was no positive sample in the original subset;
- for "Court," we randomly selected 67, 10, and 10 positive samples from "Selection" subset and added them into "Train", "Validation" and "Fusion" subset, respectively;
- for "Prisoner," we randomly selected 19 positive samples from "Fusion" subset and added them into "Validation" subset.

To address the question that "Do TRECVID 2005 positive samples facilitate the concept detection," we also built the detectors only using TRECVID 2007 data for the concepts in Table 4, except for "Airplane." We call these detectors as "TRECV07Model." Considering the performance over "Selection" subset, the TRECV0705Model performs better than TRECV07 for several concepts (as highlighted in Table 4). For these concepts and "Flag-US," we adopted the corresponding TRECV0705Model as their baseline models.

Table 4: The number of auxiliary positive samples from TRECVID 2005 development set. TRECV0705Model performs better than TRECV07 in the highlighted concepts.

Concept	# of auxiliary positive samples from TRECVID 2005
Airplane	77
Bus	24
Charts	33
Desert	97
Explosion_Fire	110
Flag-US	124
Military	45
Mountain	85
Natural-Disaster	76
People-Marching	112
Snow	36
Sports	53
Truck	81

3.2.3 Parameter Selection

The effectiveness of SVM classifiers is highly subject to the selection of model parameters. Since we use RBF kernels, we have to tune two parameters: C (the cost parameter in soft-margin SVMs) and γ (the width of the RBF function). To address the unbalance problem, we set different cost parameters for positive samples and negative samples. Thus, we should consider three model parameters: C^+ (the cost parameter for the positive examples), C^- (the cost parameter for the negative examples), and γ. In practical implementation, we assigned the ratio $\frac{C^+}{C^-}$ as $\frac{N^+}{N^-}$, where N^+ is the number of positive training examples and N^- is that of the negative training examples. Base on the "Validation" subset described in section 3.1.1, we selected the best choice of these parameters. The principle is the Average Precision (AP) on the "Validation" subset.

3.2.4 Modality Fusion

After we have learned separated models for each visual feature, we applied the late fusion approach to combine all detection results generated by different features shown in Figure 3. As reported in [5] the late fusion (i.e., combination of separated scores) performs favorably compared to early fusion (i.e., the combination of many features into one large feature vector). In particular, we evaluated three fusion strategies, including **Linear, Average, and Max.**

- **Linear:** We performed a grid search in fusion parameter space to select the optimal modality weights.
- **Average:** The scores resulting from each modality were simply averaged to generate the fused score.
- **Max:** For each concept, we selected the modality gains the best performance as the final classifier. Its results were considered as the fused results. The principle for modality selection is the Average Precision (AP) [1] performance on the "Fusion" subset.

Based on the "Selection" subset, we find linear weighted fusion performs better than the other two methods. Thus linear weighted fusion is adopted for baseline model construction.

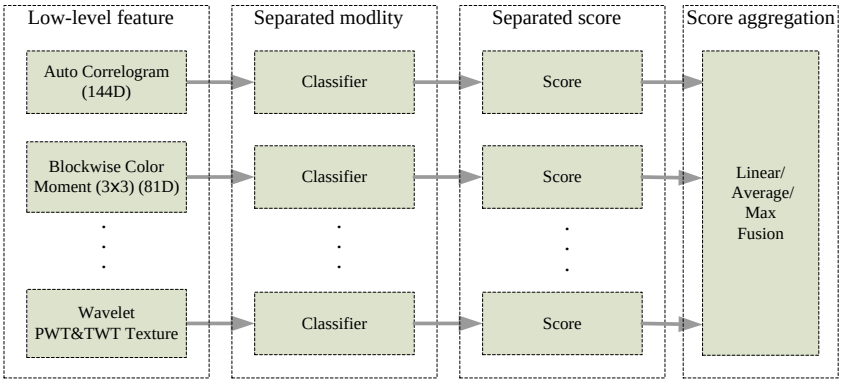

Figure 3: Late fusion of multiple modalities.

3.3 Computational complexity

We used 35 computers with dual CPUs to train all the SVM classifiers. This process took approximately two months of machine time (single CPU time) in total. For each modality, the average machine time is obtained by simply averaging the machine times for learning this modality over all the 36 concepts. Figure 4 shows these average machine times and the basic unit is "hour." The detailed machine time for each concept can be found in Appendix B.

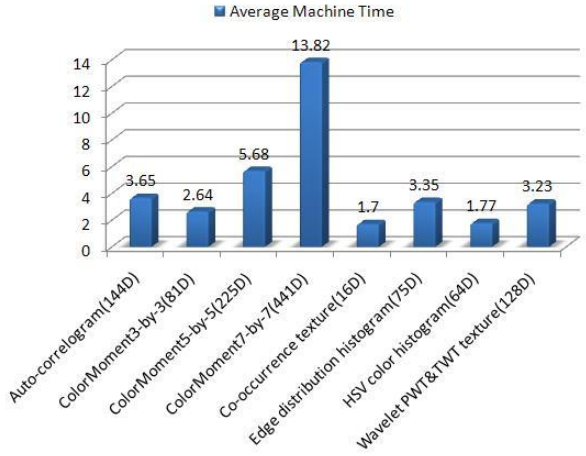

Figure 4: The Average Machine Time for learning each modality (in hour).

3.4 Performance Evaluation and Applications

We evaluated the concept detectors on the "Selection" subset. Figure 5 illustrate the detailed APs of all the 36 concepts. The Mean Average Precision (MAP) is around 0.286.

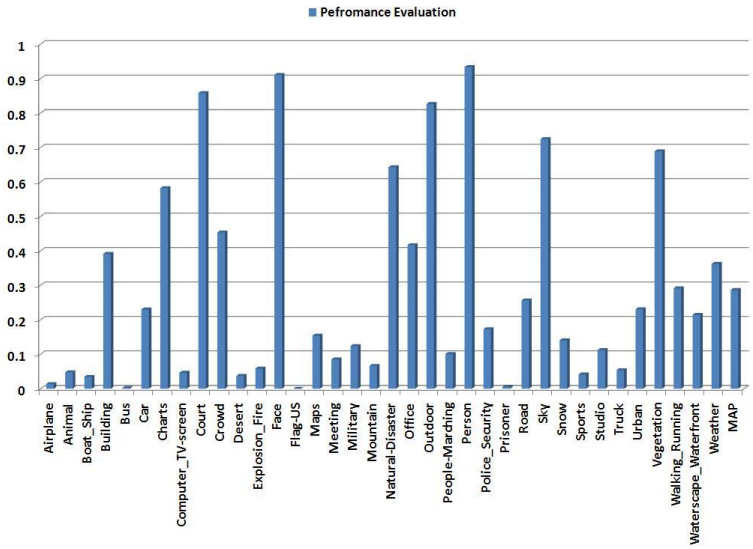

Figure 5: Performance evaluation on "Selection" subset.

Furthermore, we applied these detectors to TRECVID 2007 test set. Combined with other learning methods such as semi-supervised learning [8], multi-layer multi-instance kernel [7], etc., we obtained competitive detection results in high level feature extraction task. Furthermore, two concept-based search methods were introduced to improve search performance in automatic search task.

4 References

[1] TRECVID. http://www-nlpir.nist.gov/projects/trecvid/.

[2] NIST. http://www.nist.gov

[3] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for Support Vector Machine," 2001.

[4] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu "Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts," Columbia University ADVENT Technical Report # 222-2006-8, March 20, 2007.

[5] B.L. Tseng, C.-Y. Lin, M. Naphade, A. Natsev, and J.R.Smith, "Normalized Classifier Fusion for Semantic Visual Concept Detection," in Proceedings of IEEE ICIP, 2003.

[6] X.-S. Hua, T. Mei, W. Lai, M. Wang, J. Tang, G.-J. Qi, L. Li and Z. Gu, "Microsoft Research Asia TRECVID 2006: High-level Feature Extraction and Rushes Exploitation," in NIST TRECVID Workshop, November 2006.

[7] Z. Gu, T. Mei, X.-S. Hua, J. Tang, and X. Wu, "Multi-layer multi-instance kernel for video concept detection," in Proceedings of ACM Multimedia, 2007.

[8] M. Wang, X.-S. Hua, Y. Song, X. Yuan, S. Li, and H.-J. Zhang, "Automatic video annotation by semi-supervised learning with kernel density estimation," in Proceedings of ACM International Conference on Multimedia, Santa Barbara, CA, Oct 2006.

Appendix A: The distribution of TRECVID 2007 samples.

Concept	Positive	Negative	Skip
Airplane	31	21486	15
Animal	671	20789	72
Boat_Ship	256	21229	47
Building	1722	19717	93
Bus	38	21485	9
Car	594	20851	87
Charts	75	21407	50
Computer_TV-screen	384	21053	95
Court	111	21402	19
Crowd	3497	18023	12
Desert	32	21467	33
Explosion_Fire	72	21408	52
Face	9241	11794	497
Flag-US	12	21519	1
Maps	125	21407	0
Meeting	308	21001	223
Military	426	21088	18
Mountain	107	21400	25
Natural-Disaster	29	21496	7
Office	1318	20025	189
Outdoor	7076	13646	810
People-Marching	238	21294	0
Person	12881	7826	825
Police_Security	195	21241	96
Prisoner	81	21433	18
Road	1396	20014	122
Sky	2294	18687	551
Snow	137	21389	6
Sports	482	21048	2
Studio	341	21181	10
Truck	151	21219	162
Urban	1669	19838	25
Vegetation	2946	18559	27
Walking_Running	1985	19412	135
Waterscape_Waterfront	949	20496	87
Weather	124	21406	2

Appendix B: Machine time for each concept (in hour)

Concept	Auto-Correlogram (144D)	Color Moment 3-by-3 (81D)	Color Moment 5-by-5 (225D)	Color Moment 7-by-7 (441D)	Co-currence texture (16D)	Edge Distribution Histogram (75D)	HSV Color Histogram (64D)	Wavelet PWT&TWT Texture (128D)	Total
Airplane	3.15	2.15	5.07	10.78	0.87	1.95	1.58	2.38	27.93
Animal	4.55	3.47	8.25	20.95	1.73	3.35	2.03	4.43	48.76
Boat_Ship	4.60	3.40	7.37	17.18	1.43	2.40	1.75	2.98	41.11
Building	5.38	4.07	8.90	22.58	1.90	3.60	2.03	6.48	54.94
Bus	3.23	2.37	5.93	11.4	1.25	2.28	1.52	2.68	30.66
Car	4.32	4.45	7.13	20.68	1.77	3.30	2.08	3.75	47.48
Charts	3.32	2.32	5.38	13.25	1.22	1.88	1.38	2.72	31.47
Computer_TV-screen	3.85	2.68	6.07	16.02	1.82	3.05	1.72	3.73	38.94
Court	3.60	1.78	4.92	28.63	0.63	3.60	1.00	2.30	46.46
Crowd	7.15	4.65	11.0	23.85	3.27	4.90	2.77	6.85	64.51
Desert	2.88	1.53	3.98	7.97	0.95	1.77	1.30	2.22	22.60
Explosion_Fire	3.63	2.70	5.80	14.10	1.57	2.35	1.95	3.07	35.17
Face	4.58	3.17	6.12	16.58	3.12	3.32	3.35	7.62	47.86
Flag-US	3.22	2.93	6.20	15.38	2.87	3.65	1.38	2.95	38.58
Maps	3.67	3.65	5.97	15.25	1.82	2.45	1.77	2.77	37.35
Meeting	3.75	3.07	5.60	15.57	1.48	2.47	1.83	2.87	36.64
Military	4.83	3.43	7.35	18.98	1.78	2.93	1.75	3.43	44.48
Mountain	3.85	2.75	5.58	12.15	1.47	2.33	1.73	3.00	32.86
Natural-Disaster	2.12	1.20	3.03	6.42	1.17	1.72	0.80	2.22	18.68
Office	2.75	1.87	4.17	7.58	1.33	1.97	1.48	2.25	23.40
Outdoor	3.03	2.17	4.42	7.67	1.80	2.45	2.02	2.70	26.26
People_Marching	2.22	1.38	3.40	6.67	0.97	1.70	1.02	1.78	19.14
Person	3.33	2.35	4.67	8.22	2.12	2.50	2.48	3.02	28.69
Police_Security	3.60	2.38	4.90	10.32	1.88	2.47	1.83	2.75	30.13
Prisoner	3.27	2.20	5.00	18.37	3.28	29.83	1.62	3.98	67.55
Road	3.85	2.68	5.57	13.30	1.92	2.78	2.12	3.17	35.39
Sky	3.25	2.20	4.85	11.20	1.75	2.13	1.80	2.92	30.10
Snow	2.63	1.82	4.55	10.03	1.18	1.80	0.90	2.15	25.06
Sports	3.32	3.18	5.83	10.43	1.78	2.38	1.57	2.78	31.27
Studio	2.63	1.68	4.32	9.13	1.10	2.00	1.08	2.10	24.04
Truck	3.43	2.38	5.10	9.68	1.30	2.18	1.82	2.72	28.61
Urban	3.87	2.75	5.90	13.27	1.97	2.85	2.13	3.22	35.96
Vegetation	3.67	2.95	5.95	12.13	1.83	3.05	2.12	3.50	35.20
Walking_Running	4.25	3.02	6.12	18.88	2.25	2.95	2.52	3.55	43.54
Waterscape_Waterfront	3.47	2.40	5.38	13.33	1.47	2.40	1.83	2.88	33.16
Weather	3.05	1.87	4.53	9.48	1.05	2.02	1.50	2.43	25.93
Average	3.65	2.64	5.68	13.82	1.70	3.35	1.77	3.23	35.83