

CU-VIREO374: Fusing Columbia374 and VIREO374 for Large Scale Semantic Concept Detection

Yu-Gang Jiang, Akira Yanagawa, Shih-Fu Chang, and Chong-Wah Ngo
{*yjiang, cwngo*}@*cs.cityu.edu.hk*; {*akira, sfchang*}@*ee.columbia.edu*

Columbia University ADVENT Technical Report #223-2008-1

Aug 24, 2008

(Download Site – <http://www.ee.columbia.edu/dvmm/CU-VIREO374>)

1 Introduction

Semantic concept detection is an active research topic as it can provide semantic filters and aid in automatic search of image and video databases. The annual NIST TRECVID video retrieval benchmarking event [1] has greatly contributed to this area by providing benchmark datasets and performing system evaluation. As acquiring ground truths of semantic concepts is time-consuming, in the TRECVID event only 10-20 concepts were selected for evaluation each year. This is insufficient for general video retrieval tasks, for which most researchers believe that hundreds or thousands of concepts would be more appropriate [2]. In light of this, several efforts have developed and released annotation data for hundreds of concepts [3, 4, 5].

Although the annotations are publicly available, building detectors for hundreds of concepts is complicated and time-consuming. To stimulate innovation of new techniques and reduce the effort in replicating similar methods, there are several efforts in developing and releasing large-scale concept detectors, including Mediamill-101 [5], Columbia374 [6], and VIREO374 [7]. The Mediamill-101 includes 101 detectors over TRECVID 2005/2006 datasets, including ground truth labels, features, and detection scores. Columbia374 and VIREO374 released detectors for a larger set of 374 semantic concepts selected from the LSCOM ontology [4]. Columbia374 employed a simple and efficient baseline method using three types of global features. VIREO374 also adopted similar framework, but with an emphasize on the use of local keypoint features.

While keypoint features describe the local structures in an image and do not contain any color information, global features are statistics about the overall distribution of color, texture, or edge information in an image. Hence, we expect these two types of features are complementary for semantic concept detection, which requires either global color information (e.g. for concepts *water*, *desert*), or local structure information (e.g., for *US-flag*, *car*), or both (e.g., for *moutain*). It is interesting not only to compare the performance of various features, but also to see whether their combination further improves the performance. As Columbia374 and VIREO-374 work on the same set of concepts, we **unify the output formats** and **fuse the detection scores** of

both detector sets. With the goal of stimulating innovation in concept detection and providing better large-scale concept detectors for video search, we are releasing the fused detection scores on TRECVID 2008 corpora to the multimedia community.

2 Fusion of Columbia374 and VIREO374

Table 1 shows the features used in both detector sets. For each feature, a SVM classifier was trained and the combination of different features is done by “average fusion”, i.e. the final score is obtained by averaging outputs of multiple individual SVM classifiers. Note that as the grid-based color moment in the two sets was calculated in different color spaces, we include both of them in the fusion process. We directly combine the six classifiers to generate the final fusion scores. For more implementation details about Columbia374 and VIREO374, please refer to [6] and [7] respectively.

Table 1: Features used in Columbia374 and VIREO374.

	Feature	Dimension
Columbia374	Grid-based Color Moment (LUV)	225
	Gabor Texture	48
	Edge Direction Histogram	73
VIREO374	Bag-of-Visual-Words (soft-weighting [8])	500
	Grid-based Color Moment (<i>Lab</i>)	225
	Grid-based Wavelet Texture	81

Both Columbia374 and VIREO374 were trained on TRECVID-2005 development set, in which all videos are broadcast news. The domain is different from that of TRECVID-2008 data (foreign documentary videos from Sound and Vision). It is well-known that the domain change will hurt the detection performance [9], though additional approaches for handling cross-domain model adaptation may be explored. In TRECVID 2007, the collaborative annotation effort has annotated 36 concepts on the 2007 development data (50 hours; cf. Figure 1). To alleviate the problem caused by domain change, we retrain detectors of the 36 concepts using the 2007 development data [9, 10], and update prediction scores of the 36 concepts over the remaining 150 hours video data (50 hours of TRECVID 2007 test set and 100 hours of TRECVID 2008 test set). Table 2 summarizes the training data used for Columbia374, VIREO374, and CU-VIREO374; detailed per-concept information can be found in Appendix.

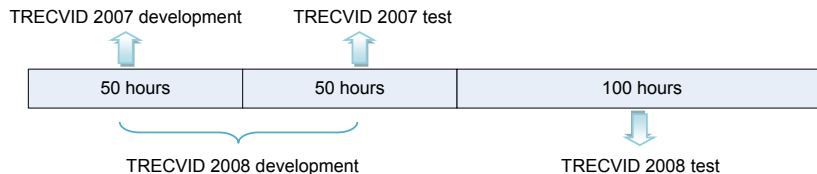


Figure 1: Data partitions in TRECVID 2007 and 2008. Both development and test data of TRECVID 2007 are used as development data of TRECVID 2008.

Table 2: Training data for Columbia374, VIREO374, and CU-VIREO374.

	Columbia374	VIREO374	CU-VIREO374
Training Data	TV'05 Devel (374)	TV'05 Devel (374)	TV'05 Devel (338); TV'07 Devel (36)

3 Performance Evaluation

We test the performance of Columbia374, VIREO374, and their fused models on both TRECVID 2006 and 2007 test data sets. For each year’s benchmark, we apply our models to the test data for the 20 concepts officially evaluated in each year. Note the 20 concepts evaluated by TRECVID 2006 are different from those evaluated in 2007. TRECVID evaluated only 20 of the 36 announced/annotated concepts in 2007. Also for 2007 test set, we apply the models that have been retrained using the 2007 development data set. The performances are shown in inferred average precision (AP), which is an approximation of conventional average precision. Inferred AP is the official metric in both TRECVID 2006 and 2007 evaluations.

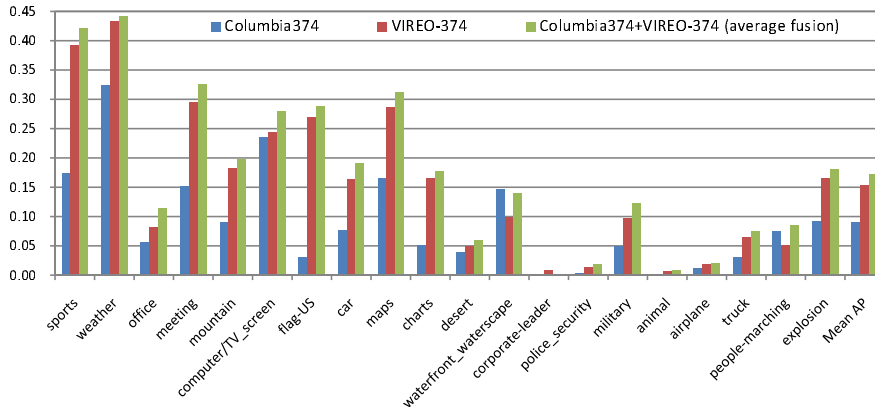


Figure 2: Per-concept analysis on TRECVID 2006 test data. The models tested here are trained using TRECVID 2005 development data.

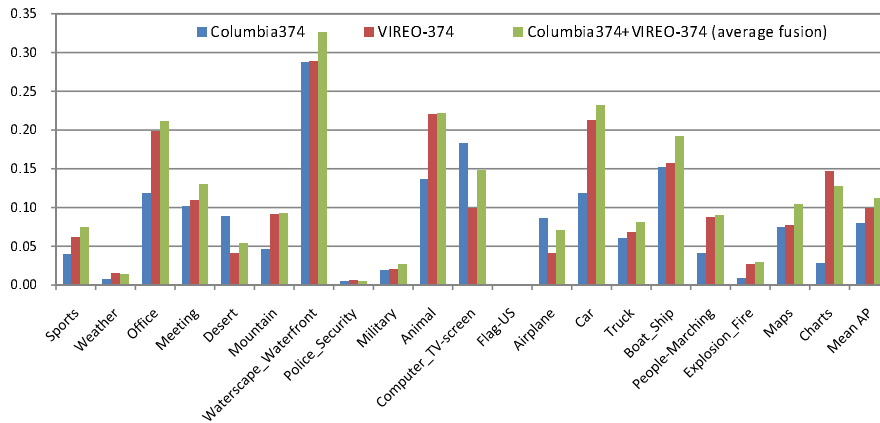


Figure 3: Per-concept analysis on TRECVID 2007 test data. The models tested here are re-trained using TRECVID 2007 development data.

Figure 2 and Figure 3 shows the per-concept comparison of Columbia374, VIREO374, and their fusion. The performances of Columbia374 and VIREO374 are obtained from the average fusion of their three feature modalities respectively. From the figures we can see that VIREO374 outperforms Columbia374 for most concepts on both benchmarks. This verifies that local key-point features are more effective for semantic concept detection, but at the price of higher computational cost for feature extraction. The fusion of Columbia374 and VIREO374 gives better or comparable performance for virtually all the concepts. In term of mean AP over the 20 concepts, the fusion performance is 0.173 on TRECVID 2006 and 0.111 on TRECVID 2007, and the improvements are respectively 12% and 14% over the higher of the two detector sets.

Figure 4 and Figure 5 show the performance comparison of Columbia374, VIREO374, and their fusion with all official concept detection systems in TRECVID 2006 and 2007 respectively. The results verify that the features from both detector sets are complementary, and our results of such a simple system are already comparable to the best few systems on both benchmarks.

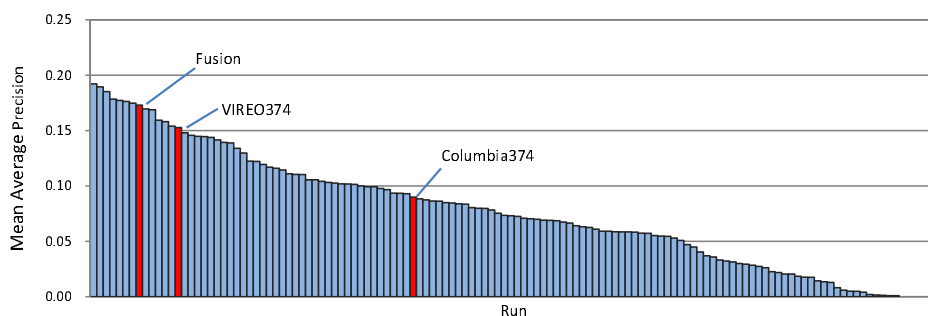


Figure 4: Performance comparison of Columbia374, VIREO374, and their fusion with all official TRECVID 2006 concept detection systems.

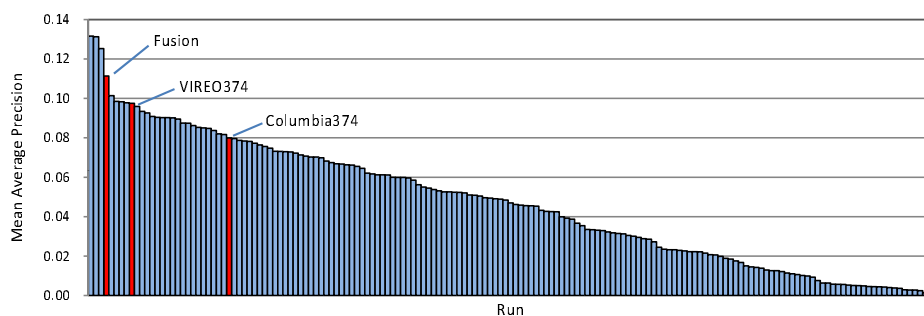


Figure 5: Performance comparison of Columbia374, VIREO374, and their fusion with all official TRECVID 2007 concept detection systems.

4 Folder Structure and File Format

In this section, we describe the file format of our detection scores on TRECVID 2008 data set, which are available for download at (<http://www.ee.columbia.edu/dvmm/CU-VIREO374>). The

features of TRECVID 2008 data can be found in the websites of [6] and [7] respectively.

There are two folders containing the detection scores over TRECVID 2008 development and test data respectively. For the development data, our detection scores are on keyframe level, where the keyframes are from the LIG group in France. For the test data, our scores are on shot level based on the shot boundaries provided by NIST. In each of the two folders, there are 374 folders for each of the concepts in our set. Note that for the 36 concepts officially announced in TRECVID 2007, we have replaced the detection scores with those generated by the re-trained models. Specifically, for the 50 hours of TRECVID 2007 development data, we replace the scores with ground-truth labels (0/1), and for the rest 150 hours video data, we use the prediction scores of the re-trained models.

Each of the 374 folders has seven score files: “*_cugcm.res”, “*_cugbr.res”, “*_cuedh.res”, “*_vireobow.res”, “*_vireogcm.res”, “*_vireogwt.res”, and “*_ave.res”, where * indicates the concept name; the first six files contain the detection scores separately using the six features shown in Table 1, and the last file includes scores from average fusion of the six feature modalities.

In each score file, all scores are listed in a column, and there are two separate files indicating the keyframe/shot names corresponding to each row of the score files, namely “list_tv08devel.txt” and “list_tv08test.txt” for TRECVID 2008 development and test data respectively.

References

- [1] A. F. Smeaton, P. Over, and W. Kraaij, “Evaluation campaigns and trecvid,” in *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*. New York, NY, USA: ACM Press, 2006, pp. 321–330.
- [2] A. Hauptmann, R. Yan, and W.-H. Lin, “How many high-level concepts will fill the semantic gap in news video retrieval?” in *Proceedings of the 6th ACM international conference on Image and video retrieval, 2007*, pp. 627–634.
- [3] “LSCOM lexicon definitions and annotations,” in *DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, Columbia University ADVENT Technical Report #217-2006-3*, 2006.
- [4] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, “Large-scale concept ontology for multimedia,” *IEEE Multimedia Magazine*, vol. 13, no. 3, 2006.
- [5] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders, “The challenge problem for automated detection of 101 semantic concepts in multimedia,” in *Proceedings of the ACM International Conference on Multimedia*, October 2006, pp. 421–430.
- [6] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu, “Columbia university’s baseline detectors for 374 lscom semantic visual concepts,” Columbia University, Tech. Rep., March 2007.
- [7] Y. G. Jiang, C. W. Ngo, and J. Yang, “VIREO-374: LSCOM semantic concept detectors using local keypoint features,” in <http://vireo.cs.cityu.edu.hk/research/vireo374/>.
- [8] Y. G. Jiang, C. W. Ngo, and J. Yang, “Towards optimal bag-of-features for object categorization and semantic video retrieval,” in *ACM CIVR*, 2007.
- [9] S. F. Chang, W. Jiang, A. Yanagawa, and E. Zavesky, “Columbia university trecvid 2007 high-level feature extraction,” in *NIST TRECVID Workshop*, 2007.
- [10] C. W. Ngo, Y. G. Jiang, X. Wei, and et. al., “Experimenting VIREO-374: Bag-of-Visual-Words and visual-based ontology for semantic video indexing and search,” in *NIST TRECVID Workshop*, 2007.

Appendix

Table 3: Training data for each concept of CU-VIREO374 models.

Concept Name	Training Data		Evaluated by NIST in	Concept Name	Training Data		Evaluated by NIST in
	05 Dev	07 Dev			05 Dev	07 Dev	
Actor	✓			Address_Or_Speech	✓		
Administrative_Assistant	✓			Adobehouses	✓		
Adult	✓			Agent	✓		
Agricultural_People	✓			Aircraft_Cabin	✓		
Airplane		✓	TV'06/07	Airplane_Flying	✓		TV'08
Airplane_Landing	✓			Airplane_Takeoff	✓		
Airport	✓			Airport_Or_Airfield	✓		
Alley	✓			Animal		✓	TV'06/07
Animal_Pens_And_Cages	✓			Antenna	✓		
Apartment_Complex	✓			Apartments	✓		
Armed_Person	✓			Armored_Vehicles	✓		
Artillery	✓			Asian_People	✓		
Athlete	✓			Attached_Body_Parts	✓		
Baby	✓			Backpack	✓		
Backpackers	✓			Baker	✓		
Bar_Pub	✓			Baseball	✓		
Basketball	✓			Bathroom	✓		
Bazaar	✓			Beach	✓		
Beards	✓			Bicycle	✓		
Bicycles	✓			Birds	✓		
Blank_Frame	✓			Boat_Ship		✓	TV'07/08
Body_Parts	✓			Bomber_Bombing	✓		
Boy	✓			Bride	✓		
Bridges	✓		TV'08	Briefcases	✓		
Building		✓	TV'05	Bus		✓	TV'08
Business_People	✓			Cables	✓		
Camera	✓			Canal	✓		
Canoe	✓			Capital	✓		
Car		✓	TV'05/06/07	Car_Crash	✓		
Car_Racing	✓			Cart_Path	✓		
Castle	✓			Caucasians	✓		
Celebration_Or_Party	✓			Celebrity_Entertainment	✓		
Cell_Phones	✓			Charts		✓	TV'06/07
Cheering	✓			Child	✓		
Cigar_Boats	✓			Cityscape	✓		TV'08
Civilian_Person	✓			Classroom	✓		TV'08
Clearing	✓			Clock_Tower	✓		
Clouds	✓			Cloverleaf	✓		
Coal_Powerplants	✓			Colin_Powell	✓		
Commentator_Or_Studio_Expert	✓			Commercial_Advertisement	✓		
Computer_Or_Television_Screens	✓			Computers	✓		
Computer_TV-screen		✓	TV'06/07	Conference_Buildings	✓		
Conference_Room	✓			Congressman	✓		
Construction_Site	✓			Construction_Vehicles	✓		
Construction_Worker	✓			Cordless	✓		
Corporate_Leader	✓			Corporate_Leader	✓		TV'06
Court		✓		Courthouse	✓		
Crowd		✓		Cul-de-sac	✓		
Dancing	✓			Dark-skinned_People	✓		
Daytime_Outdoor	✓			Dead_Bodies	✓		
Demonstration_Or_Protest	✓		TV'08	Desert		✓	TV'06/07
Dining_Room	✓			Dirt_Gravel_Road	✓		
Ditch	✓			Dogs	✓		TV'08
Donald_Rumsfeld	✓			Dredge_Powershovel_Dragline	✓		
Dresses	✓			Dresses_Of_Women	✓		
Driver	✓		TV'08	Earthquake	✓		
Election_Campaign	✓			Election_Campaign_Address	✓		
Election_Campaign_Convention	✓			Election_Campaign_Debate	✓		
Election_Campaign_Greeting	✓			Emergency_Medical_Resp_People	✓		
Emergency_Room	✓			Emergency_Vehicles	✓		TV'08
Entertainment	✓			Exiting_Car	✓		
Exploding_Ordinance	✓			Explosion_Fire		✓	TV'05/06/07
Eyewitness	✓			Face		✓	

Concept Name	Training Data		Evaluated by NIST in	Concept Name	Training Data		Evaluated by NIST in
	05 Dev	07 Dev			05 Dev	07 Dev	
Factory	✓			Factory_Worker	✓		
Farms	✓			Female_Anchor	✓		
Female_News_Subject	✓			Female_Person	✓		
Female_Reporter	✓			Fields	✓		
Fighter_Combat	✓			Finance_Busines	✓		
Firefighter	✓			First_Lady	✓		
Flags	✓			Flag-US		✓	TV'05/06/07
Flood	✓			Flowers	✓		TV'08
Flying_Objects	✓			Food	✓		
Football	✓			Forest	✓		
Foxhole	✓			Free_Standing_Structures	✓		
Freighter	✓			Funeral	✓		
Furniture	✓			Gas_Station	✓		
George_Bush	✓			Girl	✓		
Glass	✓			Glasses	✓		
Golf	✓			Golf_Course	✓		
Golf_Player	✓			Government_Leader	✓		
Government-Leader	✓			Grandstands_Bleachers	✓		
Grassland	✓			Graveyard	✓		
Greeting	✓			Groom	✓		
Ground_Combat	✓			Ground_Crew	✓		
Ground_Vehicles	✓			Group	✓		
Guard	✓			Guest	✓		
Gym	✓			Hand	✓		TV'08
Handshaking	✓			Harbors	✓		TV'08
Head_And_Shoulder	✓			Head_Of_State	✓		
Helicopter_Hovering	✓			Helicopters	✓		
High_Security_Facility	✓			Highway	✓		
Hill	✓			Horse	✓		
Hospital	✓			Host	✓		
Hotel	✓			House	✓		
House_Of_Worship	✓			Hu_Jintao	✓		
Individual	✓			Indoor_Sports_Venue	✓		
Industrial_Setting	✓			Infants	✓		
Insurgents	✓			Interview_On_Location	✓		
Interview_Sequences	✓			Islands	✓		
John_Edwards	✓			John_Kerry	✓		
Judge	✓			Kitchen	✓		TV'08
Laboratory	✓			Lakes	✓		
Landlines	✓			Landscape	✓		
Laundry_Room	✓			Lawn	✓		
Lawyer	✓			Logos_Full_Screen	✓		
Machine_Guns	✓			Male_Anchor	✓		
Male_News_Subject	✓			Male_Person	✓		
Male_Reporter	✓			Maps		✓	TV'05/06/07
Medical_Personnel	✓			Meeting		✓	TV'06/07
Microphones	✓			Military		✓	TV'06/07
Military_Base	✓			Military_Buildings	✓		
Military_Personnel	✓			Moonlight	✓		
Mosques	✓			Motorcycle	✓		
Mountain		✓	TV'05/06/07/08	Muddy_Scenes	✓		
Mug	✓			Muslims	✓		
Natural-Disaster		✓		Natural_Disasters	✓		
Network_Logo	✓			Newspapers	✓		
News_Studio	✓			Nighttime	✓		TV'08
Non-uniformed_Fighters	✓			Non-us_National_Flags	✓		
Observation_Tower	✓			Oceans	✓		
Office		✓	TV'06/07	Office_Building	✓		
Officers	✓			Oil_Drilling_Site	✓		
Oil_Field	✓			Old_People	✓		
Outdoor		✓		Outer_Space	✓		

Concept Name	Training Data		Evaluated by NIST in	Concept Name	Training Data		Evaluated by NIST in
	05 Dev	07 Dev			05 Dev	07 Dev	
Overlaid_Text	✓			Parade	✓		
Parking_Lot	✓			Pavilions	✓		
Peacekeepers	✓			Pedestrian_Zone	✓		
People_Crying	✓			People_Marching	✓		
People_Marching		✓	TV'06/07	Person		✓	
Photographers	✓			Pickup_Truck	✓		
Pipes	✓			Police	✓		
Police_Private_Security_Personnel	✓			Police_Security		✓	TV'06/07
Politics	✓			Powerlines	✓		
Power_Plant	✓			Powerplants	✓		
Power_Transmission_Line_Tower	✓			Press_Conference	✓		
Prisoner		✓	TV'05	Processing_Plant	✓		
Protesters	✓			Radar	✓		
Raft	✓			Railroad	✓		
Rainy	✓			Religious_Figures	✓		
Reporters	✓			Residential_Buildings	✓		
Rifles	✓			Riot	✓		
River	✓			River_Bank	✓		
Road		✓		Road_Block	✓		
Road_Overpass	✓			Rocky_Ground	✓		
Room	✓			Rowboat	✓		
Rpg	✓			Ruins	✓		
Running	✓			Runway	✓		
Scene_Text	✓			School	✓		
Science_Technology	✓			Security_Checkpoint	✓		
Ship	✓			Shooting	✓		
Shopping_Mall	✓			Sidewalks	✓		
Singing	✓		TV'08	Single_Family_Homes	✓		
Single_Person	✓			Single_Person_Female	✓		
Single_Person_Male	✓			Sitting	✓		
Sketches	✓			Sky		✓	
Smoke	✓			Smoke_Stack	✓		
Snow		✓		Soccer	✓		
Soldiers	✓			Speaker_At_Podium	✓		
Speaking_To_Camera	✓			Sports		✓	TV'05/06/07
Stadium	✓			Standing	✓		
Steeple	✓			Still_Image	✓		
Stock_Market	✓			Store	✓		
Street_Battle	✓			Streets	✓		TV'08
Striking_People	✓			Studio		✓	
Studio_With_Anchorperson	✓			Suburban	✓		
Suits	✓			Sunglasses	✓		
Sunny	✓			Supermarket	✓		
Swimmer	✓			Swimming	✓		
Swimming_Pools	✓			Talking	✓		
Tanks	✓			Telephones	✓		TV'08
Television_Tower	✓			Tennis	✓		
Tent	✓			Text_Labeling_People	✓		
Text_On_Artificial_Background	✓			Throwing	✓		
Ties	✓			Tony_Blair	✓		
Tower	✓			Traffic	✓		
Trees	✓			Tropical_Settings	✓		
Truck		✓	TV'06/07	Tunnel	✓		
Underwater	✓			Urban		✓	
Urban_Park	✓			Urban_Scenes	✓		
Us_Flags	✓			Valleys	✓		
Vegetation		✓		Vehicle	✓		
Walking	✓			Walking_Running		✓	TV'05
Warehouse	✓			Waterscape_Waterfront		✓	TV'05/06/07
Water_Tower	✓			Waterways	✓		
Weapons	✓			Weather		✓	TV'06/07
White_House	✓			Windows	✓		
Windy	✓			Yasser_Arafat	✓		